

Intelligenza Artificiale al Bivio: Large Language Models e World Models a Confronto

Introduzione generale

L'ora della biforcazione

Negli anni che separano la pubblicazione di *Attention Is All You Need* (Vaswani et al., 2017) dal momento in cui il presente paper viene composto, l'intelligenza artificiale ha attraversato una trasformazione che non trova precedenti nella storia della disciplina per rapidità e ampiezza delle conseguenze epistemiche. In un arco di tempo straordinariamente breve rispetto alla geologia delle rivoluzioni scientifiche, sistemi che operano su sequenze di simboli discreti hanno raggiunto livelli di competenza su compiti linguistici e inferenziali che soltanto qualche anno prima sembravano accessibili soltanto alla cognizione umana adulta. Il pre-addestramento massivo su corpus di scala planetaria, la messa a punto mediante feedback umano, l'emergere di capacità impreviste oltre soglie di scala — questi elementi hanno prodotto una generazione di modelli che influenzano la pratica professionale, la ricerca scientifica e le conversazioni pubbliche in modo che nessuno dei loro architetti aveva anticipato nella pienezza delle sue implicazioni.

Eppure, al culmine di questo successo straordinario, la disciplina è pervasa da una sensazione di incompiutezza teorica e di disomogeneità strutturale che non è riconducibile a carenze tecniche ordinarie, ma tocca questioni di principio sull'architettura della cognizione artificiale. Ilya Sutskever e Andrei Karpathy, tra i principali architetti del paradigma dominante, hanno entrambi suggerito pubblicamente che ci si stia avvicinando ai rendimenti decrescenti dell'approccio di scaling dei modelli linguistici, identificando questioni che non sono risolvibili con più dati e più parametri: l'incapacità di acquisire conoscenza genuinamente nuova al di là del corpus di addestramento, l'eccessivo logocentrismo rispetto al ragionamento spaziale e incarnato, l'assenza di un meccanismo endogeno di apprendimento continuo dopo il dispiegamento, la persistente confusione tra plausibilità statistica e veridicità fattuale. Questi non sono deficit accidentali eliminabili per affinamento: emergono dalla struttura stessa del paradigma, dal fatto che un sistema

addestrato a massimizzare la verosimiglianza del token successivo su corpus testuali non ha, nelle sue fondamenta computazionali, alcun meccanismo per distinguere la rappresentazione del mondo dalla rappresentazione della rappresentazione del mondo.

Parallelamente a questa traiettoria — e in dialogo critico con essa — si è sviluppato un programma di ricerca alternativo, o meglio complementare, che ha trovato espressione tecnica nelle Joint Embedding Predictive Architectures (JEPA) e articolazione teorica più ampia nel framework dell'apprendimento autonomo proposto da Yann LeCun nel documento seminale del 2022 *A Path Towards Autonomous Machine Intelligence*. Questo programma parte da un'intuizione radicalmente diversa sul problema dell'intelligenza: non la distribuzione statistica del linguaggio scritto ma la previsione semantica nello spazio delle rappresentazioni latenti, non la reattività autoregressiva ma la pianificazione deliberativa mediata da un modello interno del mondo, non l'accumulo passivo di correlazioni testuale ma l'acquisizione attiva di conoscenza attraverso l'interazione interventistica con l'ambiente.

L'orizzonte di questa biforcazione si precisa ulteriormente nella lettura combinata di due documenti pubblicati tra la fine del 2025 e il marzo del 2026 — ossia nei mesi immediatamente antecedenti alla composizione del presente lavoro — che costituiscono, insieme alla monografia di cui questa Introduzione fa parte, le fonti primarie di riferimento per il confronto che ci proponiamo. Il primo è il tutorial sistematico sulle JEPA curato da Monemi, Chinipardaz, Rasti, Bennis e Latva-aho del Centre for Wireless Communications dell'Università di Oulu (dicembre 2025), che offre la prima rassegna comprensiva e formalmente rigorosa delle architetture JEPA, delle loro derivazioni e delle loro applicazioni attraverso modalità sensoriali diverse, dai domini spaziali a quelli temporali, dalle reti neurali convoluzionali alle comunicazioni wireless di sesta generazione. Il secondo è il documento di Emmanuel Dupoux, Yann LeCun e Jitendra Malik intitolato *Why AI Systems Don't Learn and What to Do About It: Lessons on Autonomous Learning from Cognitive Science* (marzo 2026), che situa il programma di ricerca JEPA nel contesto del problema molto più vasto dell'apprendimento autonomo in organismi biologici e sistemi artificiali, proponendo un'architettura integrata — il framework Sistema A-B-M — che aspira a riconciliare apprendimento per osservazione, apprendimento per azione e meta-controllo cognitivo in un sistema unificato capace di apprendere come gli organismi biologici.

La monografia che qui si introduce non è un resoconto neutro di questi sviluppi. È un tentativo di mettere a fuoco, con la precisione che soltanto l'analisi formale e l'argomentazione filosofica permettono, la natura del bivio che il titolo annuncia: un bivio che non è soltanto tra due architetture o due

paradigmi computazionali, ma tra due concezioni dell'intelligenza stessa, due modelli della conoscenza, due teorie implicite di cosa significhi comprendere.

Il problema dell'apprendimento autonomo come orizzonte regolativo

La distinzione che Dupoux, LeCun e Malik pongono al centro della loro analisi — tra sistemi che apprendono e sistemi che operano su conoscenza pre-acquisita — illumina retroattivamente l'intera storia dell'intelligenza artificiale come oscillazione tra due poli che raramente sono stati tenuti insieme in modo architetturealmente coerente. Da un lato, la tradizione che parte dalla statistica, dalla teoria dell'informazione e dall'apprendimento automatico supervisionato: un sistema impara da un corpus di dati etichettati o non etichettati, ottimizza un obiettivo ben definito, e poi opera su casi nuovi senza modificare il proprio repertorio di comportamenti. Dall'altro, la tradizione che parte dal controllo, dalla cibernetica e dal rinforzo: un agente interagisce con un ambiente, riceve segnali di feedback e aggiusta la propria politica in funzione dell'esperienza accumulata.

Queste due tradizioni hanno prodotto sub-discipline separate, con terminologie incompatibili, criteri di valutazione diversi e comunità di ricercatori che si frequentano soltanto obliquamente. I modelli linguistici di larga scala appartengono interamente alla prima tradizione nella loro forma standard: sono sistemi che apprendono da un corpus immoto, che non interagiscono con il mondo durante l'addestramento se non attraverso la mediazione testuale dell'esperienza umana codificata, e che una volta dispiegati non acquisiscono più nulla che non sia loro esplicitamente somministrato attraverso procedure di fine-tuning — procedure che, per essere efficaci, richiedono l'intervento di team di esperti umani nella selezione dei dati, nella progettazione della perdita e nell'adattamento dell'obiettivo di ottimizzazione.

Il sistema A-B-M di Dupoux et al. aspira a superare questa frattura non attraverso una sintesi eclettica, bensì attraverso un'integrazione architettonica in cui l'apprendimento per osservazione (Sistema A), l'apprendimento per azione (Sistema B) e il meta-controllo che orchestra la transizione tra le due modalità (Sistema M) coesistono in un sistema unificato capace di selezionare autonomamente la propria strategia di apprendimento in funzione di segnali interni di competenza, incertezza e novità. Questa integrazione non è soltanto una proposta ingegneristica: è una tesi sulla natura dell'intelligenza, che trova fondamento nella psicologia dello sviluppo (Gopnik, Piaget, Vygotsky), nella neuroscienza cognitiva (Rao e Ballard, Friston, Miller e Cohen), e nella filosofia dell'azione (Gibson, Merleau-Ponty).

Le JEPA entrano in questo quadro come il meccanismo fondamentale del Sistema A: un modo di apprendere rappresentazioni semanticamente ricche dall'osservazione senza richiedere etichette esterne, basato sulla previsione delle rappresentazioni latenti di parti non osservate di un campione a partire dalle sue parti osservate. L'eleganza concettuale di JEPA risiede nel fatto che questo obiettivo di previsione latente — la minimizzazione di $d(\hat{s}_y, s_y)$ dove $\hat{s}_y$ è la previsione della rappresentazione del bersaglio e s_y è la sua rappresentazione effettiva — produce, quasi come corollario, le proprietà semantiche e le capacità di pianificazione che un sistema intelligente richiede: il filtro delle ridondanze irrilevanti, l'astrazione delle strutture invarianti, la costruzione di uno spazio latente in cui la prossimità metrica corrisponde alla somiglianza semantica piuttosto che alla somiglianza percettiva superficiale.

Struttura e propositi del saggio

Il presente contributo è organizzato in tre parti che procedono dal particolare al generale, dall'analisi tecnica all'elaborazione filosofica, ma che in ogni momento cercano di mantenere il filo che connette la precisione matematica alle questioni di principio che essa illumina senza esaurire.

La **Prima Parte** — *Fondamenti, Architetture e Formalizzazioni* — analizza in dettaglio i due paradigmi nella loro struttura interna. La Sezione I è dedicata ai Large Language Models: dalla derivazione analitica del meccanismo di self-attention scalata, passando per la modellazione del linguaggio come ottimizzazione della cross-entropia sulla previsione del token successivo, fino alle scaling laws di Kaplan et al. e alla revisione di Hoffmann et al., alla fenomenologia delle capacità emergenti, alla procedura di allineamento attraverso Reinforcement Learning from Human Feedback e Direct Preference Optimization. L'accento cade non soltanto sulla fenomenologia delle prestazioni ma sull'analisi delle limitazioni strutturali: il problema del grounding simbolico, la dipendenza dalla finestra contestuale, la modalità dell'allucinazione fattuale come conseguenza della struttura dell'obiettivo di addestramento, il deficit di pianificazione deliberativa come conseguenza della natura autoregressiva della generazione.

La Sezione II è dedicata ai World Models e all'architettura JEPA nella sua forma teorica e nelle sue concretizzazioni empiriche. Il punto di partenza è la genealogia intellettuale dei modelli del mondo — da Craik alle reti predittive di Rao e Ballard, dal principio di energia libera variazionale di Friston agli sviluppi del reinforcement learning model-based — e la sua confluenza con il programma JEPA inaugurato da LeCun. Il quadro formale è sviluppato nella sua completezza: la formulazione energetica dell'obiettivo di previsione, la struttura del predittore condizionato sulle variabili ausiliarie, la

regolarizzazione anti-collasso nella doppia forma degli schemi teacher-student con EMA e dei termini VICReg espliciti, le estensioni gerarchiche e sequenziali che abilitano il ragionamento multi-risoluzione e il planning multi-passo. A questa trattazione di base si articola una rassegna sistematica delle varianti architetturali documentate nella letteratura — I-JEPA, MIM-JEPA, CNN-JEPA, C-JEPA, SToP-JEPA, D-JEPA, IWM, 3D-JEPA, Point-JEPA, T-JEPA, TS-JEPA, Audio-JEPA, A-JEPA, MC-JEPA, V-JEPA, V-JEPA 2, TI-JEPA, M3-JEPA, JEP-KD, LLM-JEPA, Graph-JEPA — organizzata per modalità sensoriale e per paradigma di estensione. Speciale attenzione è riservata alla proprietà di stima implicita della densità che emerge dall'obiettivo JEPA, formalizzata attraverso il JEPA-SCORE di Balestriero et al. (2025), e alle applicazioni nelle comunicazioni wireless di sesta generazione, dalla comunicazione semantica al channel charting, dall'ISAC multimodale all'allocazione adattiva delle risorse radio.

Il framework A-B-M di Dupoux, LeCun e Malik viene trattato come l'orizzonte sistemico all'interno del quale JEPA acquista il suo significato più pieno: non soltanto come tecnica di apprendimento della rappresentazione, ma come componente di un'architettura cognitiva integrata che aspira a riprodurre — in forma artificiale e parziale — la plasticità adattiva degli organismi biologici. Il quadro di ottimizzazione bi-livello che il framework propone per il bootstrap del sistema è analizzato nella sua struttura matematica, nelle sue affinità con il meta-learning e con la neural architecture search, e nelle sue sfide di scalabilità che rimangono aperte.

La **Seconda Parte** — *Anatomia di un Confronto: Convergenze Strutturali, Divergenze Fondamentali e il Problema dell'Apprendimento Autonomo* — conduce il confronto diretto tra i due paradigmi lungo dimensioni tecniche e concettuali. L'analisi comparativa copre le epistemologie implicite, la struttura degli obiettivi di addestramento e le loro implicazioni geometriche, l'efficienza campionaria e la dipendenza dalla scala, la capacità di pianificazione deliberativa e le sue formalizzazioni, il problema del grounding fisico nella prospettiva del symbol grounding problem di Harnad e della critica searleana, le modalità qualitative dell'errore — allucinazione fattuale versus errore di previsione — come specchi delle strutture profonde dei rispettivi obiettivi, la sistematicità compositiva nella dimensione linguistica e nella dimensione fisica, il problema della causalità come gerarchia a tre livelli di Pearl e il posizionamento dei due paradigmi rispetto ad essa.

La **Terza Parte** — *Approfondimenti Formali: Causalità, Meta-Controllo e Convergenza dei Paradigmi* — approfondisce tre nodi teorici che la parte precedente individua ma non esaurisce. Il primo è la relazione tra previsione statistica e inferenza causale, sviluppata attraverso l'apparato formale del do-calculus di Pearl, il problema dell'identificabilità causale, l'analisi della competenza pseudo-causale degli LLM e della via dell'intervento incarnato

come accesso parziale al secondo livello della gerarchia causale, e il programma di Causal Representation Learning di Schölkopf come framework per valutare in che misura JEPA apprenda variabili causalmente rilevanti. Il secondo è il Sistema M come architettura meta-cognitiva, sviluppato nelle sue relazioni con la prefrontal cortex biologica, nelle sue realizzazioni parziali nell'IA attuale (active learning, prioritized experience replay, curriculum learning, hierarchical reinforcement learning) e nella sua proposta come control plane integrato per l'apprendimento autonomo. Il terzo è il problema della convergenza tra LLMs e World Models come il programma di ricerca dominante del prossimo decennio, esaminato nella sua plausibilità tecnica, nei suoi prerequisiti teorici irrisolti e nelle sue implicazioni per la questione filosofica della comprensione genuina.

Nella sua parte finale, il paper affronta le dimensioni filosofiche del confronto — la fenomenologia della comprensione, l'argomento della stanza cinese, il problema della coscienza nella prospettiva della Integrated Information Theory e della Global Workspace Theory, il problema dell'intersoggettività come condizione della comprensione genuina — e le questioni etiche specifiche sollevate dai sistemi di apprendimento autonomo: il problema dell'allineamento dinamico in sistemi che si modificano, il trade-off tra adattabilità e controllabilità, la questione della responsabilità morale in sistemi che generano comportamenti emergenti, lo statuto morale di agenti che elaborano segnali somatici funzionalmente analoghi al dolore, la sovranità cognitiva come problema di governance della fitness function evolutiva.

Una nota sul metodo

Il lettore noterà che questa monografia fa uso simultaneo di notazione matematica formale, analisi filosofica, riferimenti alla neuroscienza cognitiva e argomentazioni di carattere epistemologico. Questa pluralità non è eclettismo: rispecchia la convinzione che le questioni fondamentali sull'intelligenza artificiale non siano accessibili da un'unica prospettiva disciplinare, e che qualsiasi trattazione che si limiti a uno solo di questi registri — o alla sola analisi tecnica, o alla sola riflessione filosofica — produca necessariamente una comprensione parziale e deformante.

La notazione matematica è impiegata con precisione ma senza feticismo: le derivazioni formali sono incluse quando illuminano aspetti della struttura che la prosa non riesce a catturare con la stessa chiarezza, non per conferire autorità retorica alle posizioni sostenute. Le formule utilizzano notazione Unicode e testo piano dove le convenzioni tipografiche standard non sono immediatamente disponibili: $\hat{y}$ per le stime latenti, $\|\cdot\|_2$ per la norma

euclidea, $\mathbb{E}$ per il valore atteso, $\otimes$ e $\oplus$ per le operazioni vettoriali. Quando un'espressione matematica appare nel testo, il lettore può assumere che l'apparato formale a cui fa riferimento sia stato sviluppato in una delle sezioni precedenti.

La scelta di adottare la lingua italiana per una monografia su un campo prevalentemente anglofono è essa stessa una posizione: rivendicare che il pensiero scientifico rigoroso non richiede la lingua franca anglosassone come vettore necessario, e che la tradizione intellettuale italiana — dall'epistemologia di Enriques e Mach, alla filosofia della mente di Bozzi e Zanforlin, alla semiotica di Eco e alla filosofia del linguaggio di Bottari — ha contributi irriducibili al dibattito sulla cognizione e sulla rappresentazione che meritano di essere messi in dialogo con le questioni più tecniche che questa monografia affronta.

Il bivio come categoria intellettuale

Il titolo scelto per questo lavoro — *Intelligenza artificiale al bivio* — potrebbe sembrare, a prima lettura, una delle tante metafore stradali che il discorso sull'IA adopera con eccessiva facilità. Il significato che vi attribuiamo è più preciso e più impegnativo.

Il bivio non è tra due tecnologie concorrenti che si disputano la preferenza del mercato: è tra due domande radicalmente diverse su cosa debba essere un sistema intelligente, e quindi su cosa significhi costruirne uno. La domanda dei modelli linguistici — nella sua formulazione più ambiziosa — è: *possiamo creare un sistema che, avendo appreso le strutture del pensiero umano dalla totalità del suo output scritto, sia in grado di generalizzare al pensiero nuovo con la stessa fluidità con cui lo fa un essere umano istruito?* La domanda dei World Models e dell'apprendimento autonomo — nella sua formulazione più ambiziosa — è: *possiamo creare un sistema che apprenda a comprendere il mondo interagendo con esso, come fa un organismo biologico che viene al mondo senza conoscenza codificata ma con le strutture cognitive necessarie per acquisirla attraverso l'esperienza?*

Entrambe le domande sono legittime e profonde. Il bivio non è tra una risposta giusta e una sbagliata: è tra due programmi di ricerca che, portati alle loro conseguenze logiche, implicano architetture, criteri di valutazione e concezioni dell'intelligenza incompatibili tra loro e non riconducibili a una sintesi superficiale. Comprendere questa incompatibilità — nei suoi fondamenti formali, nelle sue implicazioni empiriche e nelle sue radici filosofiche — è il proposito principale di questo paper.

Se alla fine il lettore troverà che la risposta più onesta non è né l'una né l'altra di queste domande, bensì qualcosa di più difficile e forse irraggiungibile che le contiene entrambe come casi speciali, allora il saggio avrà assolto al suo compito più importante: non risolvere il problema, ma mostrarne la struttura con una chiarezza sufficiente da rendere il prossimo tentativo di risolverlo più preciso e più consapevole.

PARTE PRIMA — Fondamenti, Architetture e Formalizzazioni

Sezione I — Il Paradigma dei Large Language Models: Fondamenti Teorici, Apparato Matematico e Limiti Strutturali

1.1 Contestualizzazione storica e genealogia intellettuale

La traiettoria che ha condotto all'attuale generazione di modelli linguistici di larga scala non è lineare né priva di fratture epistemiche. Le sue radici affondano nella statistica bayesiana del linguaggio, nell'ipotesi distribuzionale di Harris e Firth — secondo cui il significato di un'espressione è determinato dai contesti in cui essa ricorre — e nel connessionismo che, a partire dagli anni Ottanta del secolo scorso, ha tentato di incarnare la competenza linguistica in reti di unità computazionali ispirate, per analogia remota, alla neurobiologia. Il passaggio dalle reti ricorrenti (RNN, LSTM, GRU) ai meccanismi di attenzione, e da questi all'architettura Transformer, rappresenta la discontinuità tecnica più rilevante degli ultimi decenni nell'elaborazione del linguaggio naturale.

Il contributo seminale di Vaswani et al. (2017), "Attention Is All You Need", ha sancito l'abbandono della ricorsività sequenziale in favore di un meccanismo puramente attentivo che consente la parallelizzazione massiva del calcolo e la cattura di dipendenze a lungo raggio senza il degrado del gradiente che affliggeva le architetture ricorrenti. Su questa infrastruttura, il pre-addestramento autoregressivo su corpus di enormi dimensioni — prima con GPT, poi con BERT e le sue numerose derivazioni, infine con i modelli della famiglia GPT-3/4, PaLM, LLaMA e Claude — ha prodotto sistemi capaci di prestazioni straordinarie su un'ampia gamma di compiti linguistici e, progressivamente, multimodali.

Ciò che distingue la fase attuale non è soltanto la capacità di queste architetture, ma la natura qualitativa delle capacità emergenti: la generalizzazione zero-shot e few-shot, il ragionamento per analogia, la generazione coerente su orizzonti testuali molto estesi. Questi fenomeni hanno riaperto dibattiti filosofici di lungo corso sulla natura della

comprensione, dell'intenzionalità e della cognizione, che verranno affrontati nella parte terza di questo lavoro.

1.2 L'architettura Transformer: derivazione analitica e struttura computazionale

Il Transformer opera su sequenze di simboli discreti — tipicamente token ottenuti da algoritmi di tokenizzazione come Byte-Pair Encoding (BPE) o SentencePiece — trasformandoli in sequenze di vettori continui attraverso un'operazione di embedding. Dato un vocabolario V di cardinalità $|V|$, ogni token $t \in V$ è rappresentato da un vettore $x \in \mathbb{R}^d$ tramite una matrice di embedding $E \in \mathbb{R}^{|V| \times d}$.

1.2.1 Codifica posizionale

Il meccanismo di attenzione, come si vedrà, è intrinsecamente invariante alla permutazione degli elementi della sequenza. Per introdurre informazione sull'ordine posizionale, Vaswani et al. propongono una codifica sinusoidale deterministica:

$$PE(pos, 2i) = \sin(pos / 10000^{(2i / d_{\text{model}})})$$

$$PE(pos, 2i+1) = \cos(pos / 10000^{(2i / d_{\text{model}})})$$

dove $pos \in \{0, 1, \dots, T-1\}$ è la posizione nella sequenza, $i \in \{0, 1, \dots, d_{\text{model}}/2 - 1\}$ è l'indice della dimensione e d_{model} è la dimensionalità dello spazio di rappresentazione. L'input effettivo al primo strato del Transformer è quindi:

$$h^0_{\text{pos}} = x_{\text{pos}} + PE(pos, \cdot) \in \mathbb{R}^d$$

La scelta di funzioni trigonometriche non è arbitraria: essa garantisce che le distanze relative tra posizioni possano essere espresse come trasformazioni lineari delle codifiche assolute, proprietà formalmente enunciabile come segue. Per qualsiasi sfasamento posizionale k , esiste una matrice lineare $M_k \in \mathbb{R}^{(d \times d)}$ tale che:

$$PE(pos + k, \cdot) = M_k \cdot PE(pos, \cdot)$$

I modelli successivi, tra cui RoPE (Rotary Position Embedding) e ALiBi, hanno proposto alternative che migliorano la generalizzazione a lunghezze di sequenza non incontrate durante l'addestramento, ma la struttura sinusoidale rimane concettualmente fondamentale per comprendere la filosofia posizionale del Transformer.

1.2.2 Il meccanismo di self-attention: derivazione e interpretazione

Il nucleo computazionale del Transformer è il meccanismo di attenzione scalata a prodotto interno. Date le rappresentazioni $H \in \mathbb{R}^{(T \times d)}$ di una

sequenza di T token, si derivano tre matrici distinte attraverso proiezioni lineari apprendibili:

$$Q = H W_Q, W_Q \in \mathbb{R}^{(d \times d_k)}$$

$$K = H W_K, W_K \in \mathbb{R}^{(d \times d_k)}$$

$$V = H W_V, W_V \in \mathbb{R}^{(d \times d_v)}$$

dove Q (query), K (key) e V (value) hanno dimensionalità rispettivamente $(T \times d_k)$, $(T \times d_k)$ e $(T \times d_v)$. Il meccanismo di attenzione scalata è:

$$\text{Attention}(Q, K, V) = \text{softmax}(Q K^T / \sqrt{d_k}) \cdot V$$

Il fattore di normalizzazione $1/\sqrt{d_k}$ è introdotto per mitigare il problema dei gradienti evanescenti in presenza di prodotti interni di grande magnitudine quando d_k è elevato: senza questa correzione, la funzione softmax saturerebbe su valori estremi, rendendo i gradienti trascurabili. La derivazione formale del problema è immediata: se le componenti di Q e K sono variabili casuali indipendenti con media zero e varianza unitaria, allora la varianza del prodotto interno $Q \cdot K$ è d_k , e la deviazione standard è $\sqrt{d_k}$. La normalizzazione riporta la varianza all'unità.

La matrice di attenzione $A = \text{softmax}(Q K^T / \sqrt{d_k}) \in \mathbb{R}^{(T \times T)}$ codifica, per ogni posizione i della sequenza, un insieme di pesi normalizzati che determinano quanto ciascuna posizione j contribuisce alla rappresentazione aggiornata di i . Questa struttura consente la cattura selettiva di dipendenze a distanza arbitraria con complessità computazionale $O(T^2 \cdot d)$, laddove le reti ricorrenti richiedevano $O(T)$ operazioni sequenziali ma con percorsi di dipendenza proporzionali alla distanza tra gli elementi.

1.2.3 Attenzione multi-testa

Il singolo meccanismo di attenzione impone che tutte le relazioni semantiche pertinenti siano catturate in un unico sottospazio proiettivo. L'attenzione multi-testa supera questa limitazione eseguendo h meccanismi di attenzione in parallelo su sottospazi distinti:

$$\text{head}_i = \text{Attention}(H W^i_Q, H W^i_K, H W^i_V)$$

con $W^i_Q \in \mathbb{R}^{(d \times d_k/h)}$, $W^i_K \in \mathbb{R}^{(d \times d_k/h)}$, $W^i_V \in \mathbb{R}^{(d \times d_v/h)}$. I risultati delle h teste vengono concatenati e riproiettati:

$$\text{MultiHead}(H) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \cdot W_O, W_O \in \mathbb{R}^{(d_v \times d)}$$

L'interpretazione funzionale è che le diverse teste possono specializzarsi su relazioni semantiche di natura distinta: dipendenze sintattiche, coreferenza, relazioni di argomento-predicato, strutture discourse-level. Studi di analisi meccanicistica (mechanistic interpretability) hanno confermato questa

specializzazione emergente, sebbene la distribuzione dei compiti tra le teste non sia fissa né uniforme tra modelli diversi.

1.2.4 Feed-forward sub-layer e normalizzazione

Ciascuno strato del Transformer include, oltre al meccanismo di attenzione, una rete feed-forward posizionalmente separabile:

$$\text{FFN}(x) = W_2 \cdot \sigma(W_1 \cdot x + b_1) + b_2$$

dove σ è tipicamente la funzione di attivazione ReLU, GeLU o SwiGLU nelle varianti moderne. Le dimensioni intermedie sono tipicamente $d_{\text{ff}} = 4 \cdot d$, con $W_1 \in \mathbb{R}^{(d_{\text{ff}} \times d)}$ e $W_2 \in \mathbb{R}^{(d \times d_{\text{ff}})}$. Il sub-layer feed-forward funge da memoria associativa chiave-valore, come dimostrato da Geva et al. (2021): le colonne di W_1 fungono da chiavi che attivano pattern di input corrispondenti, mentre W_2 agisce come una memoria di valori. In questa prospettiva, una parte significativa della conoscenza fattuale degli LLM risiede in questi strati piuttosto che nelle matrici di attenzione.

La normalizzazione di layer (LayerNorm) è applicata prima (pre-norm) o dopo (post-norm) ciascun sub-layer:

$$\text{LayerNorm}(x) = \gamma \odot (x - \mu) / \sqrt{(\sigma^2 + \epsilon)} + \beta$$

dove μ e σ^2 sono media e varianza calcolate sull'intera dimensione del vettore x , γ e β sono parametri apprendibili di scala e traslazione, e ϵ è una costante di stabilità numerica. La normalizzazione pre-residuale (pre-norm) è oggi prevalente poiché migliora la stabilità dell'addestramento per modelli molto profondi.

La struttura completa di uno strato Transformer con connessioni residuali è quindi:

$$h' = \text{LayerNorm}(h + \text{MultiHead}(h))$$

$$h'' = \text{LayerNorm}(h' + \text{FFN}(h'))$$

Le connessioni residuali — ereditate da ResNet — consentono il passaggio diretto del gradiente attraverso la rete, mitigando il problema del gradiente che svanisce in reti molto profonde.

1.3 Obiettivo di addestramento: modellazione del linguaggio autoregressiva

L'obiettivo di pre-addestramento standard per i modelli linguistici generativi (GPT-class) è la minimizzazione della cross-entropia sulla previsione del token successivo, formulabile come massimizzazione della verosimiglianza del corpus C :

$$L(\theta) = - E_{\{x \sim D\}} [\sum_{t=1}^T \log P_{\theta}(x_t | x_1, \dots, x_{t-1})]$$

dove D è la distribuzione empirica del corpus, T è la lunghezza della sequenza e P_{θ} è la distribuzione di probabilità sul vocabolario parametrizzata dai pesi θ del modello. L'obiettivo equivale a minimizzare la divergenza di Kullback-Leibler tra la distribuzione empirica del corpus e la distribuzione modellata:

$$L(\theta) = H(p_{\text{data}}, p_{\theta}) = H(p_{\text{data}}) + D_{\text{KL}}(p_{\text{data}} || p_{\theta})$$

dove $H(p_{\text{data}})$ è l'entropia del processo generativo dei dati, che non dipende da θ e costituisce quindi un limite inferiore della perdita ottimale. La minimizzazione di $L(\theta)$ è equivalente alla minimizzazione di $D_{\text{KL}}(p_{\text{data}} || p_{\theta})$, ovvero all'avvicinamento della distribuzione del modello a quella dei dati.

La perplexity, misura standard delle prestazioni su compiti di modellazione del linguaggio, è definita come l'esponenziale della cross-entropia media per token:

$$\text{PPL} = \exp(L(\theta) / T)$$

Essa misura il numero di scelte equiprobabili che il modello deve mediamente affrontare ad ogni passo, ed è inversamente proporzionale alla confidenza con cui il modello assegna probabilità al token corretto.

1.4 Scaling laws: fondamenti empirici e implicazioni teoriche

Una delle scoperte più rilevanti dell'era dei grandi modelli è che le prestazioni degli LLM seguono leggi di potenza prevedibili rispetto a tre variabili fondamentali: il numero N di parametri del modello, la quantità D di token di addestramento e il budget computazionale C espresso in FLOP.

Kaplan et al. (2020) hanno dimostrato empiricamente che la perdita di test L segue:

$$L(N) \approx (N_c / N)^{\alpha_N} + L_{\infty}$$

$$L(D) \approx (D_c / D)^{\alpha_D} + L_{\infty}$$

con esponenti $\alpha_N \approx 0.076$ e $\alpha_D \approx 0.095$, e costanti N_c , D_c che dipendono dall'architettura e dal dominio dei dati. La combinazione delle due leggi dà luogo a un compromesso ottimale per un budget di calcolo C fisso:

$$N_{\text{opt}} \propto C^{(0.73)}, D_{\text{opt}} \propto C^{(0.27)}$$

suggerendo che, a parità di computazione, sia preferibile investire in modelli più grandi addestrati su meno dati. La revisione operata da Hoffmann et al. (2022) con il modello Chinchilla ha tuttavia corretto questa stima,

mostrando che il regime ottimale di addestramento richiede approssimativamente $D \approx 20 \cdot N$ token — ovvero che i modelli allora disponibili erano significativamente sotto-addestrati rispetto ai dati ottimali. La legge di Chinchilla è parametrizzabile come:

$$L(N, D) = E + A / N^\alpha + B / D^\beta$$

con $E = 1.69$, $A = 406.4$, $B = 410.7$, $\alpha = 0.34$, $\beta = 0.28$ (valori stimati da Hoffmann et al. su corpus filtrato). La condizione di ottimalità per il compute $C = 6 \cdot N \cdot D$ (contando multiply-accumulate come due FLOP) porta a:

$$N_{\text{opt}}(C) \propto C^{\alpha/(\alpha+\beta)} \text{ e } D_{\text{opt}}(C) \propto C^{\beta/(\alpha+\beta)}$$

che con $\alpha \approx \beta$ implica una crescita simmetrica di parametri e dati. Queste relazioni hanno implicazioni profonde non solo pratiche ma anche epistemiche: suggeriscono l'esistenza di regolarità statistiche stabili nel processo di acquisizione della conoscenza da corpus, indipendenti dai dettagli architetturali e dall'abilità specifica considerata.

1.5 Capacità emergenti e il problema della prevedibilità

Un fenomeno che ha attirato considerevole attenzione teorica è quello delle cosiddette capacità emergenti (emergent abilities), descritto da Wei et al. (2022): abilità che appaiono improvvisamente, con discontinuità apparente, al superamento di determinate soglie dimensionali, e che non erano prevedibili dall'estrapolazione delle scaling laws su scale inferiori. Esempi documentati includono l'aritmetica multi-passo, il ragionamento per analogia simbolica, la traduzione zero-shot tra lingue assenti dal pre-addestramento.

La natura di queste emergenze è dibattuta. Una linea interpretativa, sostenuta da Schaeffer et al. (2023), argomenta che l'apparente discontinuità è un artefatto delle metriche di valutazione non lineari: quando si adottano metriche continue, le capacità crescono in modo regolare anche attraverso le presunte soglie. Un'altra prospettiva, difesa da autori come Ganguli et al., sostiene che l'emergenza riflette genuini salti qualitativi nel processo di rappresentazione interna, analoghi alle transizioni di fase nei sistemi fisici critici.

Questo dibattito ha implicazioni dirette sulla questione filosofica della comprensione, che verrà esaminata nella parte terza: se le capacità emergono da una semplice estensione quantitativa, ciò suggerisce che la comprensione potrebbe essere una proprietà graduale piuttosto che binaria.

1.6 Reinforcement Learning from Human Feedback e allineamento

I modelli base pre-addestrati su previsione del token successivo non sono direttamente adatti all'interazione con utenti umani: tendono a seguire la distribuzione del corpus di addestramento piuttosto che rispondere alle intenzioni dell'utente. La tecnica del Reinforcement Learning from Human Feedback (RLHF), formalizzata da Christiano et al. (2017) e applicata su larga scala da OpenAI con InstructGPT, risolve questo disallineamento in tre fasi.

Nella prima fase, un modello di reward $R_\psi : X \rightarrow \mathbb{R}$ viene addestrato su confronti umani tra coppie di risposte (y_1, y_2) per lo stesso prompt x , ottimizzando:

$$L_R(\psi) = - E_{\{(x, y_w, y_l) \sim D\}} [\log \sigma(R_\psi(x, y_w) - R_\psi(x, y_l))]]$$

dove y_w e y_l sono rispettivamente la risposta preferita e quella non preferita dall'annotatore, e σ è la funzione sigmoide. Nella seconda fase, il modello linguistico viene ottimizzato contro il modello di reward attraverso algoritmi di policy gradient come PPO (Proximal Policy Optimization):

$$\max_{\theta} E_{\{x \sim D, y \sim \pi_{\theta}(\cdot | x)\}} [R_\psi(x, y)] - \lambda \cdot D_{KL}(\pi_{\theta}(\cdot | x) || \pi_{ref}(\cdot | x))$$

dove π_{ref} è la policy di riferimento (il modello pre-addestrato) e il termine di KL penalty impedisce deviazioni eccessive che potrebbero portare al cosiddetto reward hacking — l'ottimizzazione di proxy non allineate con le reali preferenze umane. Varianti più recenti come DPO (Direct Preference Optimization) eliminano il modello di reward esplicito, ottimizzando direttamente sulle preferenze:

$$L_{DPO}(\theta) = - E_{\{(x, y_w, y_l)\}} [\log \sigma(\beta \cdot \log(\pi_{\theta}(y_w | x) / \pi_{ref}(y_w | x)) - \beta \cdot \log(\pi_{\theta}(y_l | x) / \pi_{ref}(y_l | x)))]]$$

1.7 Limiti strutturali degli LLM: un'analisi tecnica e concettuale

L'architettura LLM presenta una serie di limitazioni che non sono accidentali ma strutturalmente iscritte nel paradigma. La loro comprensione è prerequisito per valutare in che misura i World Models con JEPa possano costituire un complemento o un'alternativa.

La prima limitazione riguarda il **grounding simbolico**. Gli LLM operano interamente nello spazio dei token, imparando associazioni statistiche tra simboli linguistici senza alcuna connessione diretta con il mondo fisico che tali simboli denotano. La distribuzione $P_{\theta}(x_t | x_{\leq t})$ è una distribuzione su simboli, non su stati del mondo. Questa condizione riproduce, a livello computazionale, il problema classico del radicamento simbolico (symbol grounding problem) formulato da Harnad (1990): i simboli ottengono

significato solo per relazione reciproca con altri simboli, in un sistema chiuso che non ha accesso diretto ai referenti extralinguistici.

La seconda limitazione è la **dipendenza dalla finestra contestuale**. Sebbene i modelli recenti abbiano esteso la lunghezza del contesto a valori molto elevati (ordine del milione di token in alcuni casi), l'attenzione quadratica $O(I^2)$ rende l'elaborazione di sequenze molto lunghe computazionalmente onerosa. Tecniche approssimate come Linear Attention, Flash Attention e varianti basate su state-space models (Mamba, H3) attenuano questo costo ma introducono compromessi tra efficienza e capacità di recupero di informazioni remote.

La terza limitazione è l'**allucinazione fattuale**, ovvero la generazione di affermazioni plausibili ma false. Questa proprietà non è un difetto implementativo correggibile ma una conseguenza della struttura dell'obiettivo di addestramento: il modello massimizza la probabilità della sequenza, non la sua veridicità. In un corpus sufficientemente grande, sequenze plausibili ma false sono frequenti quanto sequenze plausibili e vere, e il modello non ha meccanismi interni per distinguerle se non attraverso segnali statistici di co-occorrenza.

La quarta limitazione è la **mancanza di un modello del mondo fisico**. Gli LLM non hanno accesso a simulatori né a rappresentazioni strutturate della causalità fisica. La loro apparente comprensione della fisica, della matematica o della logica è mediata interamente dalla distribuzione testuale degli argomenti di questi domini nel corpus. Questo significa che le loro competenze sono altamente sensibili alla distribuzione degli esempi di addestramento e degradano sistematicamente quando i problemi presentano strutture non rappresentate nel corpus.

La quinta limitazione — e forse la più rilevante dal punto di vista della teoria dell'intelligenza — riguarda la **pianificazione e il ragionamento deliberativo**. Il meccanismo autoregressivo genera token in sequenza, in modo sostanzialmente unidirezionale, senza possibilità di valutare esplicitamente traiettorie alternative prima di impegnarsi in una scelta. Tecniche come Chain-of-Thought Prompting, Tree of Thought e MCTS (Monte Carlo Tree Search) applicato a LLM parzialmente attenuano questo limite, ma non lo risolvono architetturalmente: richiedono sampling multiplo e valutazione esterna, che non è integrata nella struttura computazionale del modello.

1.8 La questione della composizionalità e della sistematicità

Una critica di matrice fodoriana (Fodor e Pylyshyn, 1988) applicata alle reti neurali sostiene che la vera comprensione linguistica richiede sistematicità e composizionalità: la capacità di produrre e interpretare espressioni nuove a partire dalla conoscenza dei loro costituenti. Questa proprietà è

strutturalmente garantita dai sistemi formali basati su grammatiche generative, ma è solo parzialmente acquisita dagli LLM attraverso l'addestramento su corpus.

La letteratura empirica è divisa: esperimenti di Lake e Baroni (2018), Hupkes et al. (2020) e altri mostrano che le reti addestrate su linguaggi artificiali generalizzano sistematicamente solo in misura limitata rispetto a ciò che la teoria compositazionale prevedrebbe. D'altro lato, la scala degli LLM moderni ha mostrato capacità composizionali considerevolmente superiori, sollevando la questione se la sistematicità sia un requisito qualitativo discreto o un continuum lungo il quale la scala si colloca in modo monotono crescente.

Questo dibattito rimane irrisolto e costituisce uno dei nodi teorici più rilevanti per qualunque bilancio dell'intelligenza artificiale contemporanea.

Sezione II — World Models e Joint Embedding Predictive Architectures: Fondamenti Teorici, Apparato Formale e Orizzonte Applicativo

2.1 Genealogia intellettuale dei modelli del mondo: dal controllo alle rappresentazioni latenti

L'idea che un sistema intelligente debba possedere una rappresentazione interna del mondo — capace non soltanto di reagire agli stimoli presenti ma di anticipare, simulare e pianificare conseguenze future — attraversa la storia della cibernetica, della neuroscienza computazionale e dell'intelligenza artificiale con una continuità che nessuna delle rivoluzioni architetturali ha mai interrotto del tutto. Craik (1943) ne formulò la premessa fondamentale: la cognizione è adattiva nella misura in cui il soggetto è capace di costruire un modello interno della realtà su cui eseguire simulazioni anticipatorie a costo computazionale inferiore all'esperienza diretta. Il cervello non è un registratore passivo di input sensoriali bensì un motore di previsione che confronta costantemente aspettative con osservazioni, aggiornando il proprio modello interno in funzione dell'errore di predizione.

Questa intuizione ha trovato formalizzazione quantitativa nel paradigma del predictive coding, elaborato da Rao e Ballard (1999) nell'ambito della corteccia visiva: la gerarchia corticale genera previsioni top-down sugli input attesi, e soltanto il residuo — l'errore di previsione — viene propagato verso i livelli superiori. La connessione con i meccanismi dell'attenzione neuronale e con i fenomeni di soppressione dell'adattamento ha reso questo quadro influente sia in neuroscienze che nella progettazione di architetture artificiali. Karl Friston ne ha esteso la portata attraverso il principio di energia libera variazionale, riconcettualizzando la percezione, l'azione e persino l'emozione come varianti di un unico imperativo di minimizzazione della sorpresa

sensoriale — intesa nel senso tecnico dell'entropia della distribuzione marginale sugli stati sensoriali.

Nell'intelligenza artificiale, la genealogia dei modelli del mondo seguì una traiettoria parallela ma in larga misura separata. Il controllo ottimale e la programmazione dinamica, da Bellman in poi, presuppongono la conoscenza esplicita del modello di transizione dello stato: $M(s_{t+1} \mid s_t, a_t)$. Quando tale modello non è noto, il reinforcement learning model-based lo stima dai dati interattivi per poi impiegarlo nella pianificazione. La distinzione tra RL model-free — che compila direttamente l'esperienza in una policy reattiva senza rappresentazione esplicita della dinamica — e RL model-based — che apprende un simulatore interno e vi ragiona sopra — riproduce, a livello ingegneristico, la tensione filosofica tra sistemi abituali e sistemi intenzionali, tra l'automatismo del riflesso condizionato e la deliberazione prospettica.

Nei primi world model neurali, tipicamente destinati a videogiochi o ambienti simlab, la previsione avveniva nello spazio delle osservazioni crude: pixel successivi, stati grezzi. PlaNet (Hafner et al., 2019), Dreamer (Hafner et al., 2020) e varianti successive hanno spostato la previsione nello spazio latente, apprendendo rappresentazioni compresse su cui proiettare le dinamiche ambientali. Questo passaggio ha consentito guadagni notevoli in efficienza campionaria, ma ha introdotto una nuova questione epistemica: quale struttura deve avere lo spazio latente perché la pianificazione al suo interno sia trasferibile a comportamenti efficaci nel mondo reale? La risposta di JEPa a questa domanda è radicale nella sua semplicità: la struttura deve essere quella imposta dalla predittività semantica, non dalla fedeltà ricostruttiva.

2.2 Il quadro formale delle Joint Embedding Predictive Architectures

LeCun (2022), nel documento seminale che ha inaugurato il programma di ricerca, ha proposto di considerare il problema della rappresentazione autosupervisionata come un problema di minimizzazione dell'energia in uno spazio di coppie (x, y) . L'idea portante è che un buon sistema di rappresentazione non deve necessariamente ricostruire il segnale di ingresso nella sua interezza — impresa dispendiosa e spesso ridondante — bensì catturare le relazioni semantiche tra parti dello stesso campione o tra campioni correlati, operando interamente nello spazio delle rappresentazioni latenti. Questa distinzione tra architetture generative e non-generative è filosoficamente centrale: una macchina che apprende a ricostruire ha implicite assunzioni sulla struttura dello spazio di input che possono divergere dalla struttura semantica rilevante; una macchina che apprende a predire nello spazio latente può imporre che la metrica dello spazio rifletta esclusivamente le invarianze semanticamente pertinenti.

Formalmente, dato un corpus di campioni semanticamente ricchi, JEPA decompone ciascun ingresso I in un insieme di $M \geq 1$ coppie contesto-bersaglio $\{ (x[i], y[i]) \}_{i=1}^M$. Il contesto $x[i]$ e il bersaglio $y[i]$ sono porzioni complementari o trasformate di I , generate da un modulo di pair generation il cui progetto costituisce una delle scelte architetturali più critiche dell'intero sistema. Nell'implementazione I-JEPA per immagini, x è un blocco visibile ampio (scala 0.85-1.0 dell'immagine) e $y[i]$ è uno dei quattro blocchi rettangolari mascherati (scala 0.15-0.2, aspect ratio campionato in $[0.75, 1.5]$), così da imporre una varietà strutturale dei bersagli che previene soluzioni degenerate.

Il contesto e il bersaglio vengono codificati attraverso due encoder distinti:

$$s_x = f_\theta(x) \in \mathbb{R}^d \quad s_y = f_{\theta'}(y) \in \mathbb{R}^d$$

dove f_θ è l'encoder del contesto con parametri θ , e $f_{\theta'}$ è l'encoder del bersaglio con parametri θ' . Nella configurazione standard, i due encoder condividono la medesima architettura ma θ' viene aggiornato come media mobile esponenziale (EMA) di θ :

$$\theta' \leftarrow \tau \theta' + (1 - \tau) \theta, \quad \tau \in [0, 1)$$

con τ inizialmente pari a 0.996 e crescente linearmente fino a 1.0 nel corso dell'addestramento. Questo schema di aggiornamento asimmetrico, mutuato da BYOL e dalle architetture teacher-student, svolge un ruolo duplice: stabilizza l'obiettivo di previsione — impedendo che il bersaglio muti troppo rapidamente rispetto alla capacità predittiva del contesto — e, tramite il meccanismo di stop-gradient implicito nell'EMA, impedisce il collasso rappresentativo verso soluzioni costanti. La motivazione teorica è che l'aggiornamento simultaneo e simmetrico di entrambi gli encoder attraverso la stessa perdita creerebbe un equilibrio degenerare: il sistema potrebbe minimizzare la perdita di previsione mappando tutti i campioni nello stesso punto del latente, la soluzione triviale di collasso a cui la letteratura si riferisce come representational collapse.

Il predictor g_φ stima la rappresentazione del bersaglio a partire da quella del contesto e da variabili ausiliarie z che codificano informazione relazionale tra i due:

$$\hat{s}_y = g_\varphi(s_x; z)$$

Le variabili ausiliarie assumono ruoli distinti in implementazioni diverse. In I-JEPA, z codifica gli indici posizionali delle patch mascherate, la scala e il rapporto di aspetto di ciascun blocco bersaglio — consentendo al predictor di localizzare spazialmente il bersaglio nel campo visivo ed effettuare un'operazione di inpainting semantico nello spazio delle rappresentazioni piuttosto che in quello dei pixel. In IWM (Image World Model, Garrido et al., 2024), z codifica il parametro della trasformazione applicata all'immagine

(rotazione, jitter cromatico, traslazione), facendo acquisire al predictor una struttura equivariante: la previsione di come la rappresentazione trasla in funzione dell'azione è precisamente il meccanismo che rende il sistema utilizzabile come modello del mondo per la pianificazione.

La funzione di perdita è definita come errore di previsione nella norma ℓ^2 nello spazio latente, mediato sulle M coppie:

$$\mathcal{L} = (1/M) \sum_{i=1}^M D(\hat{s}_y[i], s_y[i]) = (1/M) \sum_{i=1}^M \|g_\varphi(s_x[i]; z[i]) - s_y[i]\|^2$$

La perdita complessiva su un batch di N campioni incorpora un termine di regolarizzazione anti-collasso $\mathcal{R}$ con peso λ :

$$\mathcal{L}_{\text{tot}} = \sum_{i=1}^N \mathbb{E}_{(x,y) \sim P(x,y)} [D(\hat{s}_y, s_y)] + \lambda \cdot \mathcal{R}(\{s_x, s_y\} \in [N])$$

dove le variabili di ottimizzazione includono i parametri θ, φ e la policy di campionamento $P(x, y)$. Il termine $\mathcal{R}$ può assumere varie forme funzionali: VICReg (Bardes et al., 2022) introduce tre componenti — invarianza (allineamento tra viste aumentate dello stesso campione), varianza (mantenimento della varianza marginale per ciascuna dimensione sopra una soglia) e covarianza (decorrelazione delle dimensioni). La perdita VICReg in forma esplicita è:

$$\mathcal{R}_{\text{VICReg}} = \lambda_v \cdot \sum_j \max(0, \gamma - \sqrt{\text{Var}(s_j)} + \epsilon) + \lambda_i \cdot (1/N) \sum_j \|s_x - s_y\|^2 + \lambda_c \cdot (1/d) \sum_{j \neq k} [\text{Cov}(s_j, s_k)]^2$$

dove s_j indica la j -esima dimensione della rappresentazione nel batch, $\text{Cov}(s)$ è la matrice di covarianza del batch, e γ è la varianza obiettivo (tipicamente 1). Il termine di varianza penalizza distribuzioni collassate verso costanti; quello di covarianza ortogonalizza le dimensioni, favorendo rappresentazioni non ridondanti. Questa perdita di regolarizzazione può essere interpretata come enforcement di un prior gaussiano isotropo sulla distribuzione marginale delle rappresentazioni nel batch — scelta non arbitraria, come dimostrato da Balestriero et al. (2025), che hanno mostrato che il punto di ottimo dell'obiettivo JEPA corrisponde a embedding gaussiani, proprietà sfruttabile per derivare stimatori non parametrici della densità dei dati senza alcun ulteriore addestramento.

2.3 Gerarchie, sequenzialità e multi-passo: H-JEPA e S-JEPA

L'architettura JEPA elementare cattura relazioni predittive tra coppie di viste di un singolo campione. Due estensioni naturali affrontano rispettivamente la multi-risoluzione e la dinamica temporale.

L'H-JEPA (Hierarchical JEPA) dispone più moduli JEPA in cascata gerarchica: il livello n riceve come contesto l'embedding $s_x[n-1]$ prodotto al livello inferiore e lo proietta in uno spazio di astrazione superiore $s_x[n]$,

producendo contemporaneamente la previsione del bersaglio $\hat{s}_y[n]$ tramite un predictor dedicato $g_\varphi^{(n)}$. Formalmente, per L livelli gerarchici:

$$s_x[n] = f_\theta^{(n)}(s_x[n-1]), \quad n = 1, \dots, L \quad \hat{s}_y[n] = g_\varphi^{(n)}(s_x[n]; z[n])$$

$$\mathcal{L}_H = \sum_{i=1}^L (1/M) \sum_i \|g_\varphi^{(n)}(s_x^{(i)}[n]; z^{(i)}[n]) - s_y^{(i)}[n]\|_2^2$$

Il livello inferiore opera su caratteristiche locali e transitorie (rilevanti per la previsione a breve termine), mentre i livelli superiori producono embedding astratti che catturano strutture semantiche invarianti a trasformazioni di bassa rilevanza (illuminazione, lieve occlusione, angolo di ripresa). Questa gerarchia d'astrazione rispecchia la struttura della corteccia visiva, dove le aree ventrali tardive (IT corticale) codificano identità degli oggetti in forma astratta rispetto alla posizione, alla scala e all'orientamento.

L'S-JEPA (Sequential JEPA) dispone invece una catena di predictor in serie, dove ciascun modulo n genera la previsione dello stato latente successivo $\hat{s}_y[n+1]$ condizionatamente all'embedding $\hat{s}_y[n]$ e a variabili latenti ausiliarie $z[n]$:

$$\hat{s}_y[n+1] = g_\varphi^{(n)}(\hat{s}_y[n]; z[n])$$

Questa struttura implementa rollout multi-step nello spazio latente, fondamentale per la pianificazione: fornito un contesto iniziale s_x e un orizzonte temporale M , il sistema simula la traiettoria delle rappresentazioni future $\{\hat{s}_y[1], \dots, \hat{s}_y[M]\}$ senza mai ricostruire le osservazioni grezze. La compatibilità con il paradigma di Model Predictive Control (MPC) è immediata: dati N candidati di traiettoria d'azione $\{\tilde{a}^{(n)}\}_{n=1}^N$, si seleziona quella che minimizza la funzione di costo aggregata $\sum C(\hat{s}^{(n)}[t+m])$.

2.4 V-JEPA e il dominio video: spaziotemporalità e previsione semantica

V-JEPA (Bardes et al., 2024) estende il paradigma al dominio video, introducendo la sfida della coerenza spaziotemporale. Un clip di N_f fotogrammi viene dapprima suddiviso in token spaziotemporali attraverso una convoluzione 3D:

$$I \in \mathbb{R}^{(N_f \times H \times W \times 3)} \rightarrow \text{tokens} \in \mathbb{R}^{(N_t \times d)}$$

dove N_t è il numero totale di token dopo il flattening lungo le dimensioni temporale e spaziale. La strategia di mascheramento adottata è il 3D Multi-Block Masking: si definisce una maschera spaziale 2D come unione di blocchi contigui, che viene poi replicata identicamente su tutti i fotogrammi del clip. Con un rapporto di mascheramento del 90%, l'encoder del contesto elabora soltanto il 10% dell'informazione video, rendendo il compito di previsione estremamente impegnativo e forzando l'acquisizione di rappresentazioni semanticamente ricche piuttosto che di semplice memorizzazione.

Il dato tecnico di maggior interesse è che V-JEPA, addestrato esclusivamente su previsione latente senza alcuna etichetta, supera modelli precedenti addestrati supervisionate su classificazione video e generalizza a classificazione su ImageNet — dimostrando che le rappresentazioni spaziotemporali semanticamente coerenti acquisite attraverso la previsione latente catturano strutture visive trasferibili trasversalmente ai domini.

V-JEPA 2 (Assran et al., 2025) introduce il condizionamento sull'azione, consentendo la pianificazione goal-conditioned. Il predictor autoregressivo a 300M parametri riceve come input la rappresentazione dello stato corrente $s_x[t]$ e le azioni tokenizzate $a[t]$ (vettori 7-dimensionali nello spazio delle giunture robotiche), producendo la previsione dello stato latente futuro:

$$\hat{s}_y[t+1] = g_\psi(s_x[t]; a[t])$$

In fase di pianificazione, si impiegano rollout latenti multi-step con ottimizzazione delle traiettorie attraverso Cross-Entropy Method (CEM): si campionano 800 traiettorie candidate nello spazio delle azioni, si valuta per ciascuna il costo della traiettoria latente simulata, e si raffina iterativamente la distribuzione di campionamento verso le elite. I risultati su compiti di manipolazione robotica zero-shot (65-80% di successo) confrontati con baseline generative mostrano che la pianificazione nello spazio latente semantico è non solo computazionalmente più efficiente rispetto alla pianificazione nello spazio dei pixel, ma anche più robusta alla variabilità irrilevante dell'input.

2.5 I-JEPA e la previsione semantica di immagini: un approfondimento tecnico

L'implementazione I-JEPA (Assran et al., 2023) merita un'analisi dettagliata perché costituisce il paradigma architetturale di riferimento e ha stabilito la validità empirica del quadro teorico su larga scala. Entrambi gli encoder — contesto e bersaglio — sono Vision Transformers (ViT), architetture introdotte da Dosovitskiy et al. (2020) che trattano un'immagine come sequenza di patch tokens. Per un'immagine 224×224 con patch 16×16 , si ottengono $14 \times 14 = 196$ token, ciascuno proiettato nello spazio di rappresentazione d-dimensionale.

L'encoder del contesto riceve le patch visibili dopo il mascheramento del bersaglio: la maschera viene applicata prima dell'encoding, evitando lo spreco computazionale di codificare tokens che verranno ignorati. L'encoder del bersaglio processa invece l'immagine intera, e le rappresentazioni corrispondenti alle patch bersaglio vengono estratte dopo l'encoding. Questa asimmetria — contesto-encoder sparso, bersaglio-encoder denso — è deliberata: si vuole che l'encoder del bersaglio produca rappresentazioni ricche dell'immagine completa, così da offrire supervisione semanticamente informativa al predictor.

Il predictor di I-JEPA è un ViT stretto (narrow ViT), con numero ridotto di strati e di teste di attenzione rispetto agli encoder principali. I tokens di input al predictor sono le rappresentazioni s_x a cui vengono concatenati i tokens ausiliari z che codificano gli indici posizionali delle patch bersaglio, la scala e il rapporto di aspetto di ciascun blocco. Il predictor deve quindi non soltanto propagare informazione dal contesto, ma localizzare spazialmente il bersaglio e stimarne la rappresentazione in modo compatibile con la distribuzione del bersaglio-encoder.

La funzione di perdita operativa è:

$$\mathcal{L}_{\text{I-JEPA}} = (1/4) \sum_{i=1}^4 (1/|\mathcal{P}_i|) \sum_{\Sigma \in \mathcal{P}_i} \|\hat{s}_{\{y_j\}^i} - s_{\{y_j\}^i}\|^2_2$$

dove $\mathcal{P}_i$ è l'insieme degli indici di patch nel blocco bersaglio i -esimo, $\hat{s}_{\{y_j\}^i}$ è la previsione del predictor e $s_{\{y_j\}^i}$ è la corrispondente rappresentazione del bersaglio-encoder. L'ottimizzazione aggiorna θ e φ per gradiente discendente, mentre θ' viene aggiornato per EMA.

Un risultato tecnico di notevole interesse è che ricostruire con un decoder generativo le rappresentazioni prodotte da I-JEPA rivela che queste catturano la struttura semantica degli oggetti — posizione, classe, orientamento — mentre sopprimono dettagli di texture e di sfondo irrilevanti. Questa proprietà distingue nettamente I-JEPA dai modelli ricostruttivi come MAE (Masked Autoencoder, He et al., 2022), che invece preservano dettagli di bassa rilevanza semantica per soddisfare l'obiettivo ricostruttivo pixel-level. La scelta di operare nel latente è quindi non soltanto computazionalmente vantaggiosa ma semanticamente selettiva.

2.6 Il sistema A-B-M: verso l'apprendimento autonomo

Dupoux, LeCun e Malik (2026) propongono un'architettura cognitiva integrata che situa JEPA e i World Models nel contesto più ampio del problema dell'apprendimento autonomo. La distinzione fondamentale è tra due modalità primarie di acquisizione della conoscenza — apprendimento per osservazione (Sistema A) e apprendimento per azione (Sistema B) — e un orchestratore meta-cognitivo (Sistema M) che regola il flusso di informazione tra le prime due.

Il Sistema A formalizza l'apprendimento osservativo attraverso il seguente schema: dati campioni $x \sim D$, un generatore di compiti G produce coppie $(x_{\text{in}}, x_{\text{tar}}) = G(x)$, e si cerca un encoder f_θ che minimizzi la perdita:

$$\theta^* = \operatorname{argmin}_\theta \mathbb{E}_{x \sim D} \mathcal{L}(f_\theta(x_{\text{in}}), x_{\text{tar}})$$

Questa formulazione astrae un'ampia famiglia di algoritmi: SSL su immagini (I-JEPA, DINO, SimCLR), modelli linguistici autoregressivi (GPT), masked language modeling (BERT). La struttura è passiva nel senso che la distribuzione D e il generatore G sono forniti esternamente — tipicamente

da ingegneri umani che curano il dataset e progettano il compito di preaddestramento.

Il Sistema B formalizza l'apprendimento per interazione. L'agente esegue azioni a_t nello stato s_t del mondo, ricevendo ricompense $r(s_t, a_t)$ ed effettuando una transizione secondo la dinamica $M(s_{t+1} | s_t, a_t)$. L'obiettivo è massimizzare il rendimento atteso scontato:

$$J(\pi) = \mathbb{E} [\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$$

con $s_{t+1} \sim M(\cdot | s_t, a_t)$ e $a_t \sim \pi(\cdot | s_t)$. Il parametro $\gamma \in [0,1)$ è il fattore di sconto che bilancia il peso delle ricompense immediate rispetto a quelle future.

Il Sistema M opera a un meta-livello: non elabora direttamente i flussi di dati ad alta banda (input sensoriali, comandi motori), bensì monitora telemetria di bassa dimensionalità — errori di previsione, segnali di incertezza, stati somatici — e produce meta-azioni che dinamicamente ricablano i flussi di dati tra A, B e la memoria episodica:

$$a_{m_{t+1}} \sim \Pi(\cdot | s_{m_t})$$

dove s_{m_t} è il meta-stato (errore epistemico, novità, sorpresa) e a_{m_t} è la meta-azione (connettere il flusso di A a B, attivare il replay in memoria, commutare dalla modalità inferenza a quella apprendimento). Questa struttura parallela alla control plane delle reti software-defined (SDN) è deliberata: il Sistema M non produce output cognitivi primari, ma governa la topologia dinamica dell'architettura cognitiva in risposta a segnali di bassa risoluzione.

Il framework evolutivo-sviluppativo (Evo/Devo) proposto per costruire tali sistemi istanzia una struttura di ottimizzazione bi-livello. I parametri meta φ — che codificano l'architettura iniziale, i bias induttivi e il programma sviluppativo — vengono ottimizzati al livello esterno (evolutivo) minimizzando una funzione di fitness $\mathcal{L}$ calcolata sull'intera traiettoria di vita dell'agente:

$$\varphi_{t+1} = \operatorname{argmin}_{\varphi_t} \mathcal{L}(A_0:A_K, B_0:B_K)$$

sotto i vincoli: $A_0, B_0, M = \operatorname{Init}(\varphi_t) \quad A_{i+1}, B_{i+1} = \operatorname{Update}(M, A_i, B_i, \operatorname{Env})$

Al livello interno (sviluppativo), i parametri di A e B si aggiornano per interazione con l'ambiente guidata da M. La funzione $\mathcal{L}$ e l'ambiente Env sono le sole componenti progettate manualmente: tutto il resto — regole di aggiornamento, filtri sui dati, curriculum — emerge dall'ottimizzazione di φ attraverso M. Il problema appartiene alla classe dell'ottimizzazione bi-livello, affrontata dalla letteratura sul meta-learning (Schmidhuber, 1987; MAML di

Finn et al., 2017) e sulla neural architecture search, ma a scale molto maggiori e con stati di mondo dinamici, il che introduce sfide di scalabilità ancora aperte.

2.7 JEPA-SCORE e la stima implicita della densità

Un risultato teorico di notevole interesse, recentemente dimostrato da Balestrieri et al. (2025), chiarisce una proprietà emergente dell'architettura JEPA che non era nel progetto originale. L'obiettivo anti-collasso di JEPA, nella sua formulazione ottimale, induce l'encoder f_θ a produrre embedding gaussiani: la distribuzione marginale delle rappresentazioni su cui il sistema è stato addestrato converge a una gaussiana isotropa. Questa proprietà implica una relazione tra la densità nello spazio di input p_X e la densità nello spazio delle rappresentazioni $p_{\{f_\theta(X)\}}$ mediata dallo jacobiano dell'encoder:

$$\log p_X(x) \approx \log p_{\{f_\theta(X)\}}(f_\theta(x)) + \sum \log \sigma(J_{\{f_\theta\}}(x))$$

dove $\sigma(J_{\{f_\theta\}}(x))$ sono i valori singolari dello jacobiano dell'encoder nel punto x . Le regioni dell'input manifold dove l'encoder comprime fortemente (piccoli valori singolari) corrispondono ad alta densità, mentre le regioni espanse corrispondono a bassa densità. Lo JEPA-SCORE è definito come:

$$\text{JEPA-SCORE}(x) = \sum_{i=1}^{\text{rank}(\Sigma_J)} \log \sigma(J_{\{f_\theta\}}(x))$$

con valori alti che segnalano campioni a bassa densità (outlier semantici) e valori bassi che identificano campioni nel nucleo della distribuzione dell'addestramento. Questa proprietà abilita applicazioni di data curation, rilevazione di anomalie e stima della densità senza alcun fine-tuning aggiuntivo — un'utilità accessoria che emerge gratuitamente dall'obiettivo di previsione latente.

2.8 Rassegna sistematica delle varianti architetturali di JEPA

L'ecosistema delle implementazioni JEPA si è articolato con notevole rapidità in un ventaglio di varianti che coprono modalità diverse, strategie di regolarizzazione alternative e domini applicativi specializzati. Una rassegna sistematica è indispensabile per fornire al lettore tecnico un quadro completo delle implementazioni esistenti, che la sezione precedente aveva soltanto adombrato.

Nel dominio dell'imaging statico, I-JEPA (Assran et al., 2023) è l'implementazione originale di riferimento, basata su ViT con mascheratura multi-blocco a scala variabile ($M=4$ blocchi bersaglio, scala 0.15–0.2, rapporto di aspetto in $[0.75, 1.5]$), encoder del target aggiornato per EMA con τ iniziale pari a 0.996. Il contributo di I-JEPA è stato quello di dimostrare che la previsione puramente semantica nello spazio latente — senza alcuna ricostruzione pixel-level — produce rappresentazioni di qualità superiore a

quelle dei modelli ricostruttivi come MAE su task di classificazione lineare e semi-supervisionata.

MIM-JEPA (Vélez-García et al., 2025) affronta la dipendenza di I-JEPA da dataset di grandi dimensioni introducendo SCOT²T, uno Sparse Convolutional Tokenizer for Transformers: uno strato convoluzionale superficiale che precede il ViT negli encoder e inietta bias spaziali locali nella tokenizzazione. Questa modifica migliora l'efficienza campionaria permettendo a MIM-JEPA di surclassare I-JEPA su dataset di dimensioni ridotte, con un pipeline che rimane architecturalmente fedele allo schema JEPA di base.

CNN-JEPA (Kalapos e Gyires-Tóth, 2024) abbandona completamente l'encoder ViT in favore di reti convoluzionali pure, tipicamente ResNet-50, sia per il contesto che per il target, con un predictor anch'esso interamente convoluzionale che gestisce il mascheramento tramite sparse convolution. Il risultato documentato è una superiore efficienza computazionale e prestazioni più elevate su benchmark small-scale come ImageNet-100, il che suggerisce che l'inductive bias convoluzionale può effettivamente complementare o sostituire l'attenzione globale del ViT in regimi di dati limitati.

C-JEPA (Mo e Tong) affronta due debolezze specifiche di I-JEPA: la tendenza residua al collasso parziale nonostante l'EMA, e la difficoltà del predictor nel modellare accuratamente le rappresentazioni medie dei patch. La soluzione adottata è l'integrazione della regolarizzazione VICReg nell'obiettivo di addestramento, con i tre termini di invarianza, varianza e covarianza che impongono distribuzioni marginali ben comportate. C-JEPA mantiene l'architettura ViT degli encoder e del predictor, dimostrando che i guadagni prestazionali osservati su ImageNet-1K — convergenza più rapida e accuratezza più elevata rispetto a I-JEPA — derivano esclusivamente dalla strategia di regolarizzazione e non da modifiche architetturali all'encoder.

SToP-JEPA (Bar et al., 2024) interviene su una limitazione più sottile di I-JEPA: la dipendenza da embedding posizionali fissi rende il modello eccessivamente condizionato dalla conoscenza esatta della posizione dei patch, il che diventa problematico quando l'informazione spaziale è incerta o assente. SToP-JEPA sostituisce gli embedding posizionali deterministici con variabili gaussiane la cui media corrisponde alla posizione nominale e la cui covarianza è appresa dall'encoder del contesto. Questa modifica minima — richiedente poche righe di codice aggiuntive — produce rappresentazioni più robuste all'incertezza spaziale e migliora le prestazioni di linear probing su ImageNet senza alcun costo computazionale aggiuntivo.

D-JEPA (Chen et al., 2025) spinge il framework JEPA verso la generazione condizionale d'immagini, aggiungendo obiettivi di diffusion e flow matching al loss di previsione latente standard, con un processo di generazione

autoregressivo che trasforma JEPA da rappresentazione pura in un generatore competitivo. D-JEPA raggiunge risultati all'avanguardia su generazione condizionale ImageNet, con FID inferiore e addestramento più rapido rispetto ai baseline ricostruttivi, e si estende naturalmente al video e all'audio, configurandosi come una base unificata per la modellazione generativa multimodale.

L'Image World Model (IWM, Garrido et al., 2024) adotta una strategia radicalmente diversa per la generazione delle viste: piuttosto che partizionare l'immagine in blocchi disgiunti, IWM genera la vista di contesto applicando una trasformazione parametrica (rotazione, traslazione, jitter cromatico) all'immagine target. Il parametro della trasformazione a diventa la variabile ausiliaria $z \equiv a$ che condiziona il predictor, inducendo una struttura equivariante: il predictor apprende come la rappresentazione si trasforma sotto trasformazioni fisiche o semantiche note. Questa proprietà è direttamente rilevante per l'impiego come world model, poiché la previsione di come il mondo cambia in risposta a un'azione è formalmente identica alla previsione di come la rappresentazione cambia sotto una trasformazione nota.

Nel dominio delle nuvole di punti tridimensionali, 3D-JEPA (Hu et al., 2024) è la prima estensione del framework JEPA a dati non su griglia. La tokenizzazione utilizza Farthest Point Sampling per selezionare M centri rappresentativi, KNN per raggruppare i vicini di ciascun centro e PointNet leggero per aggregare i patch in embedding. La strategia di mascheratura seleziona i centri dei blocchi bersaglio via FPS dall'insieme di contesto, riducendo la sovrapposizione e migliorando l'efficienza. 3D-JEPA introduce un decoder context-aware che alimenta continuamente le rappresentazioni di contesto codificate nei propri strati durante la previsione dei bersagli, conseguendo miglioramenti medi del 31.43% di accuratezza sui tre split di ScanObjectNN con soli 150 epoche di pre-addestramento contro le 300 tipicamente necessarie ai metodi precedenti.

Point-JEPA (Saito et al., 2025) introduce il Greedy Sequencer, un meccanismo che ordina i patch di punti vicini in sequenze coerenti, consentendo al modello di catturare la continuità semantica tra regioni locali. L'architettura mantiene encoder transformer per i flussi di contesto e di target e un predictor che allinea i loro spazi di feature. Il risultato è una convergenza più rapida e prestazioni all'avanguardia sui task di point cloud con dati etichettati scarsi.

T-JEPA (Thimonier et al., 2024) adatta il framework JEPA ai dati tabulari strutturati, un dominio dove le augmentation per SSL sono difficili da progettare senza produrre campioni irrealistici. T-JEPA tokenizza ciascuna colonna indipendentemente (normalizzazione delle feature numeriche, embedding delle categoriche, più embedding di indice e di tipo), applica

mascheratura a livello di colonna e addestra un FT Transformer backbone a predire le colonne bersaglio dalle colonne di contesto. La valutazione su benchmark di classificazione e regressione tabulare documenta rappresentazioni più ricche e rilevanti per i task rispetto ai modelli tradizionali, permettendo ai modelli deep di competere con gradient boosted decision trees.

Nel dominio audio, Audio-JEPA (Tuncay et al., 2025) estende I-JEPA all'audio trattando lo spettrogramma Mel come un'immagine mono-canale, potenzialmente non quadrata (assi tempo-frequenza). Il backbone encoder è un ViT semplice: l'encoder di contesto elabora i patch visibili mentre un encoder target aggiornato per EMA processa lo spettrogramma completo incluse le regioni mascherate. Un predictor leggero allinea le rappresentazioni di contesto e di target nello spazio latente. Audio-JEPA dimostra che questo obiettivo di mask-prediction produce embedding audio di alta qualità con dati di pre-addestramento notevolmente ridotti.

A-JEPA (Fei et al., 2023) condivide l'approccio spettrogramma-come-immagine di Audio-JEPA ma introduce un curriculum di mascheratura in due fasi: mascheratura a blocchi casuali (relativamente semplice per il modello) seguita da mascheratura strutturata tempo-frequenza che rende il task predittivo più impegnativo. Questo curriculum progressivo consente alla rete di imparare prima da contesti più semplici per poi adattarsi a pattern più complessi. Nella fase di fine-tuning, A-JEPA applica Regularized Patch Masking sull'encoder di contesto per robustezza contro l'overfitting. A-JEPA scala efficacemente e raggiunge prestazioni all'avanguardia su AudioSet-2M.

MC-JEPA (Bardes, Ponce e LeCun, 2023) affronta l'integrazione di stima del flusso ottico e apprendimento auto-supervisionato dei contenuti all'interno di un encoder condiviso. Costruendo sull'architettura del flow estimator PWC-Net e sul framework VICReg, il modello apprende simultaneamente le corrispondenze spaziali tra frame video consecutivi e le rappresentazioni semantiche del contenuto dell'immagine. Questo setup multi-task crea una sinergia tra l'obiettivo di stima del flusso e quello di SSL, producendo feature visive che incorporano sia cue di aspetto che di moto.

Nel dominio multimodale, TI-JEPA (Vo et al., 2024) costruisce uno spazio di embedding congiunto per l'allineamento testo-immagine: le immagini target e di contesto vengono prima codificate dagli encoder visivi, il testo corrispondente da un encoder testuale, e i flussi risultanti vengono integrati attraverso blocchi di cross-attention che allineano informazione testuale e visiva in uno spazio condiviso. Soltanto i moduli di cross-attention e i layer di predizione vengono fine-tunati, con gli encoder pre-addestrati mantenuti congelati per preservare la qualità delle rappresentazioni. TI-JEPA supera i

lavori all'avanguardia correlati in termini di accuratezza e F1-score mantenendo la generalizzazione multi-task.

M3-JEPA (Lei et al., 2025) è progettato per l'apprendimento multimodale any-to-any genuino: non è vincolato a una specifica coppia di modalità ed è capace di allineare qualsiasi combinazione di modalità — testo con immagine, immagine con audio, audio con testo e oltre. Al suo nucleo vi è un predittore Multi-Gate Mixture of Experts (MMoE) che crea percorsi separati per l'informazione specifica di ciascuna modalità e per quella condivisa. L'addestramento utilizza Alternating Gradient Descent, che aggiorna il predictor un task alla volta evitando i conflitti di gradiente tipici dell'apprendimento multi-task. M3-JEPA consegue risultati competitivi su diversi benchmark multimodali mostrando robustezza su dataset non visti durante il training.

JEP-KD (Sun, Qin e Yang, 2024) reinterpreta il meccanismo predittore-perdita di JEPA per il contesto cross-modale del lip reading. La modalità audio viene processata da un encoder pre-addestrato per il riconoscimento vocale, quella video da un encoder video dedicato. La feature video codificata viene combinata con un vettore di rumore casuale z e passata a un Generatore che svolge il ruolo del predictor JEPA, producendo una feature simile all'audio semanticamente arricchita. Un Discriminatore nel framework GAN a minimi quadrati sostituisce la semplice perdita basata sulla distanza di JEPA, giudicando attivamente se l'input sia una vera feature audio dall'encoder ASR teacher oppure una generata. L'addestramento segue un curriculum in tre fasi (warm-up, predizione avversariale, raffinamento), e i risultati mostrano miglioramenti rispetto ai modelli VSR baseline.

Nel dominio dei grafi, Graph-JEPA (Skenderi et al., 2025) è la prima applicazione del framework JEPA all'apprendimento di rappresentazioni a livello di grafo. Il modello allinea sottografi di contesto e bersaglio nello spazio di embedding, producendo rappresentazioni altamente semantiche per la classificazione e il ragionamento su grafi. La distinzione rispetto agli approcci GNN tradizionali sta nell'obiettivo auto-supervisionato: invece di ricostruire struttura o attributi dei nodi, Graph-JEPA impara a predire la rappresentazione di sottografi bersaglio da quella di sottografi di contesto, trasferendo al dominio discreto dei grafi la filosofia predittiva latente originariamente sviluppata per l'imaging.

DMT-JEPA (Mo e Yun, 2024) affronta una debolezza documentata di I-JEPA: la limitata capacità di catturare semantica locale fine-grained. Invece di predire patch mascherati indipendentemente, DMT-JEPA costruisce bersagli discriminativi latenti aggregando feature da patch vicini semanticamente correlati attraverso cross-attention leggera. Questa strategia di masked semantic neighboring arricchisce la comprensione locale. MaskJEPA (Kim et al., 2024) integra JEPA con architetture di mask

classification per abilitare SSL direttamente su modelli di segmentazione, usando il decoder transformer dell'architettura di mask classification come predictor nel framework JEPA. SparseJEPA (Hartman e Varshney, 2025) estende il framework aggiungendo una penalità sparsity-inducing che raggruppa variabili latenti semanticamente correlate, riducendo la ridondanza nello spazio di embedding.

2.9 JEPA per le comunicazioni wireless e le reti 6G

Un'intera frontiera applicativa di JEPA riguarda le reti di comunicazione wireless e i sistemi 6G, un dominio dove l'eleganza del paradigma predittivo latente risolve problemi di compressione e trasmissione semantica che i sistemi convenzionali non sanno affrontare.

La comunicazione semantica reinterpreta l'obiettivo della trasmissione: non inviare bit con fedeltà pixel-level o sample-level, bensì trasmettere il significato rilevante per i task a valle. TS-JEPA (Girgis, Valcarce e Bennis, 2025) incarna questa filosofia nel controllo remoto in rete: i dispositivi codificano i propri stati ad alta dimensionalità in embedding compatti tramite l'encoder di contesto JEPA e trasmettono soltanto questi embedding compatti. Il controller centrale, dotato del predictor JEPA, stima gli stati futuri a partire dagli embedding ricevuti e dai comandi di controllo previsti, senza necessità di trasmissione continua di stati grezzi. In valutazione su dati di serie temporali d'immagine da una simulazione di carrello invertito, TS-JEPA ha ottenuto una riduzione sostanziale del carico di comunicazione preservando accuratezza predittiva ed efficacia del controllo elevate.

W-JEPA (Bou Chaaya, Girgis e Bennis, 2025) estende il channel charting con JEPA per abilitare la previsione delle dinamiche canale future nello spazio latente. Il channel charting comprime la matrice CSI (Channel State Information) ad alta dimensionalità in uno spazio 2D di pseudo-localizzazione dell'utente. W-JEPA augmenta questo schema con la struttura predittiva di JEPA: l'encoder apprende la mappatura di charting, il predictor cattura le dinamiche temporali degli stati latenti guidate dalla mobilità dell'utente. Il risultato è la previsione di embedding futuri senza ricostruzione della CSI grezza, con una riduzione dei costi computazionali sostanziale. Esperimenti su dataset CSI reali mostrano che W-JEPA consegue un'accuratezza predittiva fino a due volte superiore ai benchmark per le dinamiche canale a lungo termine.

Il paradigma TokCom (Qiao et al., 2025) è un framework semantico generativo basato su tokenizzazione e codebook insieme a un foundation model generativo (GFM) o modello multimodale di linguaggio di grandi dimensioni (MLLM). Al trasmettitore, l'input di ciascuna modalità viene tokenizzato in elementi specifici di un codebook; una policy network seleziona un sottoinsieme di token da trasmettere mascherando il resto

secondo criteri di rilevanza semantica e qualità del canale. Al ricevitore, un GFM/MLLM ricostruisce i token mascherati o corrotti tramite cross-modal attention, e i detokenizzatori ricostruiscono il segnale per ciascuna modalità. JEPA offre una via complementare di integrazione con TokCom: nella fase ricevente, TokCom predice prima i token di contesto mascherati, e i token di contesto completati vengono poi passati al predictor JEPA per inferire i token bersaglio, ottenendo la rappresentazione completa con due livelli di compressione gerarchica.

Nell'ambito dell'ISAC (Integrated Sensing and Communication), Nishio, Chen e Bennis (2025) sfruttano JEPA per spostare l'integrazione e l'elaborazione multimodale dal dominio dei dati allo spazio semantico. Ciascun sensore — fotocamera, dispositivo Wi-Fi, LiDAR, radar — codifica la propria osservazione in un embedding latente compatto attraverso un encoder semantico e lo trasmette alla stazione base. Quando una modalità è assente o corrotta, la stazione base esegue semantic inpainting, predicendo gli embedding mancanti da quelli disponibili. Gli esperimenti condotti su un dataset multimodale di immagini di fotocamere e CSI Wi-Fi confermano l'efficacia del framework in termini di accuratezza e riduzione dell'overhead di trasmissione.

In un'altra direzione, Bou Chaaya, Girgis e Bennis (2025) propongono un'architettura con due JEPA accoppiati — Control JEPA e Wireless JEPA — per modellare e predire congiuntamente nello spazio latente la dinamica del sistema di controllo (rappresentata da pixel) e l'ambiente di propagazione wireless (dalla CSI). Un agente di deep reinforcement learning deriva la policy di controllo dalle dinamiche latenti di controllo, mentre un predittore di potenza di trasmissione stima gli intervalli di scheduling sulla base della rappresentazione CSI latente. I risultati su dati multimodali sintetici mostrano una riduzione della potenza trasmessa del 50% mantenendo le prestazioni di controllo.

2.10 JEPA come modello del mondo per la pianificazione agente: la visione di LeCun

LeCun (2022), nel documento seminale che ha inaugurato il programma di ricerca, distingue due modalità operative di un sistema percettivo-agente ispirandosi alla distinzione di Kahneman tra Sistema 1 e Sistema 2. La Modalità 1 (Mode-1) è reattiva: il percetto $x[t]$ viene codificato in una rappresentazione $s[t] = f_{\theta}(x[t])$ e questa viene passata direttamente a una policy π_{ψ} che produce l'azione $a[t] = \pi_{\psi}(s[t])$. Nessuna simulazione futura è eseguita; l'agente reagisce senza pianificare.

La Modalità 2 (Mode-2) introduce il world model come mediatore della pianificazione deliberativa. L'agente propone N traiettorie di azioni candidate $\hat{a}^{(n)} = (\hat{a}^{(n)}[t], \hat{a}^{(n)}[t+1], \dots, \hat{a}^{(n)}[t+M-1])$ per orizzonti di pianificazione

di M passi. Per ciascuna traiettoria, il world model simula la sequenza di rappresentazioni risultanti. Un modulo di valutazione del costo C assegna un punteggio a ciascuna traiettoria, e l'agente seleziona e ottimizza la traiettoria a costo minimo, applicando la prima azione della traiettoria ottimale all'ambiente.

In questa architettura, JEPA funge da modello di transizione interno: il predictor g_φ , addestrato in modo auto-supervisionato, apprende le dinamiche latenti del mondo e consente la simulazione di traiettorie future senza interazione diretta con l'ambiente reale. V-JEPA 2 realizza concretamente questa visione: pre-addestrato in modo auto-supervisionato su oltre un milione di ore di video, la sua variante action-conditioned (V-JEPA 2-AC) viene post-addestrata su circa 62 ore di dati di interazione robotica, e la pianificazione è eseguita tramite Model Predictive Control con Cross-Entropy Method, campionando 800 traiettorie su orizzonti da 1 a 10 passi. Il risultato — 65-80% di successo zero-shot nella manipolazione robotica — costituisce la realizzazione empirica più avanzata del Mode-2 nella letteratura JEPA.

PARTE SECONDA — Anatomia di un Confronto: Convergenze Strutturali, Divergenze Fondamentali e il Problema dell'Apprendimento Autonomo

Sezione III — Analisi Comparativa: Large Language Models e World Models a Confronto tra Teoria, Prassi e Filosofia

3.1 Due paradigmi, due epistemologie

La contrapposizione tra Large Language Models e World Models non è semplicemente una questione di scelte architetturali alternative: riflette due concezioni profondamente diverse di cosa significhi rappresentare la conoscenza, come debba essere acquisita e a quale scopo debba servire. Comprendere questa divergenza nella sua radicalità è il prerequisito per qualsiasi valutazione tecnica non superficiale.

Gli LLM incarnano ciò che potremmo chiamare un'epistemologia della prossimità distribuzionale: il significato di un'espressione — e per estensione la conoscenza del mondo che essa veicola — è interamente codificato nelle relazioni statistiche di co-occorrenza tra simboli nel corpus. La distribuzione $P_\theta(x_t \mid x_{\{<t\}})$ è una distribuzione sul vocabolario, non sugli stati del mondo; la comprensione è, in questo quadro, un effetto collaterale dell'ottimizzazione statistica. La fonte di ispirazione intellettuale più prossima è l'ipotesi distribuzionale di Firth e Harris, radicalizzata dalla disponibilità di

corpus di dimensioni senza precedenti e dalla capacità delle reti neurali di astrarre pattern a molteplici livelli di generalità.

I World Models, e in particolare l'architettura JEPA nel contesto del programma A-B-M, incarnano un'epistemologia della previsione incarnata: la conoscenza è essenzialmente capacità di simulare le conseguenze delle proprie azioni nel mondo, di anticipare la struttura degli stati futuri e di valutare traiettorie alternative prima di impegnarsi in una scelta. Il riferimento intellettuale è qui duplice: da un lato la tradizione della robotica e del controllo ottimale, dall'altro la psicologia evolutiva e la neuroscienza cognitiva che descrivono l'intelligenza come soluzione adattiva al problema della sopravvivenza in ambienti non stazionari. Craik, Gibson, Merleau-Ponty e la tradizione della cognizione incarnata costituiscono l'interlocutore filosofico implicito di questa prospettiva.

3.2 Obiettivi di addestramento a confronto: struttura formale e implicazioni cognitive

La differenza tra i due paradigmi emerge con precisione analitica confrontando le loro rispettive funzioni obiettivo. Per gli LLM autoregressivi:

$$\mathcal{L}_{\{\text{LLM}\}}(\theta) = -\mathbb{E}_{\{x \sim D\}} [\sum \log P_{\theta}(x_t | x_1, \dots, x_{t-1})]$$

L'obiettivo è una distribuzione condizionale su simboli discreti. Non vi è alcun vincolo esplicito sulla struttura geometrica dello spazio delle rappresentazioni interne — soltanto quello imposto implicitamente dalla necessità di catturare l'informazione contestuale rilevante per la previsione del token successivo. Le rappresentazioni intermedie emergono come conseguenza dell'ottimizzazione e non possono essere direttamente impiegate per pianificare sequenze d'azione o simulare l'evoluzione di stati del mondo fisico.

Per JEPA:

$$\mathcal{L}_{\{\text{JEPA}\}}(\theta, \varphi) = (1/M) \sum_i \|g_{\varphi}(f_{\theta}(x[i]); z[i]) - f_{\theta'}(y[i])\|_{2_2}^2 + \lambda \cdot \mathcal{R}(\text{embeddings})$$

L'obiettivo è una metrica nello spazio latente che misura la distanza tra la previsione e la rappresentazione effettiva del bersaglio. La struttura geometrica dello spazio latente è direttamente condizionata dall'obiettivo: le trasformazioni che non influiscono sulla predittività del contesto rispetto al bersaglio tendono a essere compresse verso l'invarianza, mentre le dimensioni predittivamente informative vengono ampliate. Lo spazio delle rappresentazioni acquisisce quindi una geometria semanticamente funzionale, non puramente statistica.

Questa differenza si cristallizza nel concetto di grounding. Le rappresentazioni degli LLM sono ancorate soltanto ad altri simboli linguistici:

la parola "tavolo" è rappresentata in relazione a "sedia", "superficie", "legno", "mangiare" — ma non all'esperienza percettiva o motoria di un tavolo. Le rappresentazioni di V-JEPA, per converso, sono apprese da sequenze video che mostrano oggetti che si muovono, collidono, vengono afferrati — e la previsione latente forza il modello ad acquisire invarianti strutturali che corrispondono, almeno parzialmente, alle categorie che la percezione umana usa per organizzare il mondo fisico. Il grounding è incompleto — manca ancora l'esperienza propriocettiva, la resistenza tattile, l'effetto dell'azione diretta — ma è qualitativamente distinto dall'assenza totale di ancoraggio fisico che caratterizza i sistemi puramente linguistici.

3.3 Efficienza campionaria e dipendenza dalla scala

Una delle critiche più incisive agli LLM, formulata con particolare acutezza nel programma di ricerca sull'apprendimento autonomo, riguarda l'efficienza campionaria. Un bambino di quattro anni ha elaborato, nella sua esperienza sensoriale accumulata, un numero di "token" di informazione enormemente inferiore a quello su cui viene addestrato un LLM di grandi dimensioni. Ciò nonostante, il bambino generalizza in modo robusto a situazioni radicalmente nuove, comprende la causalità fisica, pianifica sequenze d'azione multi-passo e apprende nuove abilità con pochissimi esempi.

La legge di scaling empirica degli LLM documenta che le prestazioni crescono secondo leggi di potenza rispetto a N (parametri), D (token di addestramento) e C (compute):

$$\mathcal{L}(N, D) = E + A/N^\alpha + B/D^\beta$$

con $\alpha \approx 0.34$, $\beta \approx 0.28$ (Hoffmann et al., 2022). Il regime di Chinchilla-optimal richiede $D \approx 20 \cdot N$ token, il che per un modello da 70B parametri implica l'addestramento su circa 1.4 trilioni di token — un dato che, traducendo in equivalente tempo di lettura umana a velocità normale, corrisponderebbe a diverse decine di migliaia di anni di lettura ininterrotta. La voracità di dati degli LLM è strutturale: l'obiettivo di previsione del token successivo non impone alcuna struttura a priori sull'informazione, dunque il modello deve indurre tutte le regolarità del mondo esclusivamente dalla co-occorrenza statistica, senza potersi appoggiare a prior cognitivi o a interazione attiva con l'ambiente.

I sistemi basati su JEPA e World Models operano in un regime di efficienza campionaria considerevolmente più favorevole per certi compiti, sebbene il confronto sia complesso perché le modalità di addestramento differiscono sostanzialmente. V-JEPA 2-AC raggiunge prestazioni significative su manipolazione robotica zero-shot dopo un fine-tuning su soli 62 ore di dati di interazione robotica — un dato che, per un sistema LLM, richiederebbe quantità enormemente maggiori di dati annotati. La ragione strutturale è che il pre-addestramento su video non annotati fornisce già rappresentazioni

spaziotemporali ricche che catturano la fisica del mondo visivo, e il condizionamento sull'azione richiede soltanto di apprendere come queste rappresentazioni trasformino in funzione di interventi agentivi — un compito molto più circoscritto che l'apprendimento da zero.

Tuttavia, questa superiorità dell'efficienza campionaria ha una contropartita: i World Models sono specializzati sulla modalità sensoriale e sul tipo di input su cui sono stati addestrati. Un modello addestrato su video di scene indoor non generalizza automaticamente a testi, conversazioni o ragionamento astratto. Gli LLM, grazie all'ampiezza del corpus linguistico, acquisiscono una forma di conoscenza enciclopedica che nessun sistema addestrato su video o sensori fisici può replicare con risorse equivalenti. La questione non è quindi quale dei due paradigmi sia superiore in senso assoluto, bensì quale tipo di generalizzazione sia strutturalmente favorita da ciascuna architettura.

3.4 Il problema della pianificazione: architetture reattive contro sistemi deliberativi

La capacità di pianificare — valutare traiettorie alternative nel futuro prima di impegnarsi in un'azione — è forse la linea di frattura più netta tra i due paradigmi. Il meccanismo autoregressivo degli LLM genera sequenze di token in modo unidirezionale e sostanzialmente irrecuperabile: ogni token viene prodotto condizionatamente alla storia precedente, senza possibilità di tornare indietro e rivalutare alternative già scartate. Tecniche come Chain-of-Thought Prompting, Tree of Thoughts e Self-Consistency attenuano questa limitazione attraverso il sampling multiplo e la valutazione di percorsi alternativi, ma questa valutazione è esterna all'architettura — non è integrata nel processo computazionale di generazione bensì è implementata come scaffolding esterno che chiama il modello più volte.

In termini formali, il meccanismo autoregressivo implementa una policy greedily one-step:

$$x \sim \operatorname{argmax}_x \{P_\theta(x \mid x_{<t})\} \quad (\text{o sampling stocastico})$$

senza la possibilità di simulare conseguenze future prima della scelta. L'equivalente cognitivista è un agente Kahneman Sistema 1 — rapido, reattivo, privo di deliberazione esplicita. Quando gli LLM vengono guidati a produrre pensieri intermedi (chain-of-thought), ciò che osserviamo è un'emulazione testuale del ragionamento deliberativo: il modello produce token che descrivono i passaggi di un ragionamento, non che eseguono il ragionamento in senso computazionale diretto.

Il paradigma World Model con pianificazione latente implementa invece un ciclo deliberativo esplicito, formalizzato nel framework Mode-2 di LeCun:

1. L'encoder estrae la rappresentazione dello stato corrente: $s[t] = f_{\theta}(x[t])$
2. L'attore propone N traiettorie candidate: $\{\tilde{a}^{(n)}\}_{n=1}^N$ con $\tilde{a}^{(n)} = (a^{(n)}[t], \dots, a^{(n)}[t+M-1])$
3. Il world model simula le rappresentazioni degli stati futuri: $s^{(n)}[t+m] = g_{\varphi}(s^{(n)}[t+m-1]; a^{(n)}[t+m-1])$
4. La funzione di costo valuta ciascuna traiettoria: $C^{(n)} = \sum_{m=0}^{M-1} C(s^{(n)}[t+m])$
5. L'attore seleziona la traiettoria ottimale $n^* = \operatorname{argmin}_n C^{(n)}$ ed esegue la prima azione

Questa architettura corrisponde a un agente Sistema 2 nel senso di Kahneman: deliberativo, multi-step, capace di valutare le conseguenze delle proprie scelte nel futuro prima di impegnarsi in un'azione irreversibile. La qualità della pianificazione è direttamente proporzionale alla fedeltà del world model nel simulare le transizioni di stato e alla ricchezza dello spazio latente nel catturare le proprietà dell'ambiente rilevanti per il compito.

La differenza tra i due approcci alla pianificazione può essere formalizzata in termini di profondità di look-ahead. Gli LLM generativi hanno look-ahead nullo per definizione architetturale (la generazione è one-token-at-a-time). I World Models implementano look-ahead di orizzonte arbitrario M , limitato soltanto dalla fedeltà del modello di transizione su traiettorie lunghe e dal budget computazionale. Il prezzo da pagare è l'accuratezza del world model: errori di previsione si accumulano lungo l'orizzonte di pianificazione secondo un regime che dipende dalla Lipschitz continuity del predictor e dalla variabilità intrinseca dell'ambiente.

3.5 Il grounding fisico e il problema della simbolicità

Il symbol grounding problem formulato da Harnad (1990) interroga la condizione di possibilità del significato in un sistema computazionale: come possono i simboli manipolati da un computer acquisire significato, se essi sono definiti soltanto per le loro relazioni con altri simboli, senza mai toccare il referente extralinguistico? L'argomento della stanza cinese di Searle (1980) radicalizza questa critica: un sistema che manipola simboli secondo regole sintattiche non possiede comprensione semantica, indipendentemente dalla correttezza e dalla complessità delle sue operazioni.

Dal punto di vista di queste critiche classiche, gli LLM sono sistemi simbolici nel senso di Harnad: operano interamente all'interno dello spazio dei token, senza accesso diretto ai referenti del mondo fisico. Le loro rappresentazioni interne sono vettori in uno spazio d -dimensionale che catturano relazioni

statistiche tra simboli, ma il collegamento con il mondo esterno è mediato interamente dalla distribuzione del corpus — una distribuzione generata da agenti umani embodied che hanno esperienza diretta del mondo, ma di cui l'LLM eredita soltanto le tracce linguistiche, non l'esperienza incarnata che le ha prodotte.

I World Models costituiscono un tentativo di affrontare — senza risolvere completamente — questo problema. JEPA addestrato su video visivi acquisisce rappresentazioni che codificano strutture geometriche, dinamiche di movimento, relazioni di occlusione e permanenza degli oggetti — invarianti percettive che la tradizione gibsoniana identificherebbe con affordances ecologicamente rilevanti. Garrido et al. (2025) hanno dimostrato che V-JEPA apprende spontaneamente forme elementari di fisica intuitiva — permanenza degli oggetti, aspettativa di continuità del moto, violazione di eventi impossibili — senza alcun addestramento esplicito su questi concetti. Questa acquisizione emerge dalla pressione selettiva della previsione latente su video naturali, che contengono le regolarità fisiche come struttura implicita della distribuzione degli stati.

Tuttavia, nemmeno i World Models visivi risolvono completamente il grounding problem. Le rappresentazioni apprese rimangono funzionali rispetto al compito di previsione video, ma non sono ancorate all'esperienza propriocettiva, alla resistenza degli oggetti alla manipolazione, al peso, alla temperatura — le dimensioni sensoriali che per un agente embodied costituiscono il substrato primario del significato fisico. Un sistema robotico che integra feedback tattile, propriocettivo e vestibolare con modelli di previsione visiva e linguistica si avvicinerebbe maggiormente a un grounding completo, ma tale integrazione rimane una sfida aperta sia tecnicamente che concettualmente.

3.6 Il dominio linguistico come asimmetria vantaggiosa

Sarebbe un errore di prospettiva presentare i World Models come superiori agli LLM senza riconoscere l'asimmetria strutturale che opera nel dominio linguistico. Il linguaggio non è soltanto un medium di comunicazione: è il supporto principale attraverso cui la conoscenza astratta — matematica, logica, filosofia, diritto, storia — viene accumulata, trasmessa e rielaborata tra generazioni. Un sistema che abbia accesso diretto a questa immensa riserva di conoscenza codificata possiede capacità di ragionamento astratto e di recupero di informazioni che nessun sistema addestrato prevalentemente su input sensoriali può replicare con risorse paragonabili.

Gli LLM sono, in questo senso, i migliori distillatori di conoscenza enciclopedica umana disponibili: il corpus su cui vengono addestrati è la più grande raccolta di pensiero umano formalizzato mai assemblata, e il processo di addestramento ne estrae strutture statistiche di ordine superiore che

corrispondono a pattern di ragionamento, analogia, inferenza e narrazione. La prova più convincente di questa capacità è la generalizzazione few-shot e zero-shot: un LLM che non ha mai visto esplicitamente un problema di geometria non-euclidea può spesso risolverlo per analogia con la geometria euclidea e con la struttura del ragionamento matematico che ha appreso dal corpus.

I World Models, nella loro configurazione attuale, sono essenzialmente ageografici rispetto al linguaggio: apprendo strutture fisiche dal mondo visivo o sensoriale, ma non hanno accesso diretto al corpus della cultura umana. L'integrazione tra world modeling e linguaggio è un'area di ricerca attiva — i Vision-Language-Action models (VLA) come RT-2 di Google tentano questa fusione — ma la combinazione non è ancora pienamente soddisfacente: i sistemi tendono a sfruttare la componente linguistica in modo superficiale, per il recupero di conoscenza dichiarativa, senza che la rappresentazione linguistica si integri con il world model a livello di rappresentazione condivisa.

3.7 Allucinazione fattuale versus errore di previsione: due fenomenologie dell'errore

Le modalità di errore dei due sistemi sono qualitativamente diverse e rivelano la struttura profonda dei rispettivi obiettivi di addestramento. Gli LLM producono allucinazioni: affermazioni fluenti, grammaticalmente corrette e semanticamente plausibili che sono fattualmente false. La causa strutturale è che l'obiettivo di massimizzazione della verosimiglianza non distingue tra proposizioni vere e proposizioni plausibili all'interno della distribuzione del corpus — in un corpus abbastanza grande, entrambe le tipologie sono rappresentate, e il modello apprende la plausibilità, non la veridicità. Il sistema non ha accesso a un oracolo esterno che possa sanzionare le sue produzioni per corrispondenza con i fatti del mondo; può soltanto fare affidamento sulla coerenza con la distribuzione interna.

I World Models commettono errori di previsione: le rappresentazioni degli stati futuri simulate dal predictor differiscono da quelle osservate. Questo tipo di errore è intrinsecamente verificabile e auto-correttivo: al termine di una sequenza di azioni pianificate, il sistema osserva lo stato reale e calcola la discrepanza rispetto alla previsione, aggiornando il modello di conseguenza. La feedback loop tra previsione e osservazione che è strutturalmente assente negli LLM — che non possono verificare le proprie produzioni confrontandole con la realtà durante la generazione — è invece il cuore architetturale dell'apprendimento attivo nei World Models. Questa proprietà non è soltanto tecnicamente rilevante: ha implicazioni per la questione filosofica della corrispondenza con la realtà che sarà approfondita nella Parte terza.

3.8 Sistematicità compositiva e generalizzazione strutturale

La critica fodoriana alla composizionalità, richiamata nella sezione 1.8 della prima parte, ha implicazioni diverse per i due paradigmi. Gli LLM acquisiscono sistematicità compositiva in misura crescente con la scala, ma questa acquisizione è statisticamente mediata: la capacità di combinare concetti in modi nuovi riflette le regolarità composizionali presenti nel corpus, non una rappresentazione esplicita delle regole generative del linguaggio. La generalizzazione è quindi robusta per combinazioni che rispettano pattern statisticamente frequenti nel corpus, ma fragile per combinazioni genuinamente nuove che violano le distribuzioni attese.

I World Models acquisiscono una forma diversa di sistematicità: la composizionalità fisica. Dato che un sistema addestrato a prevedere stati del mondo fisico apprende implicitamente le leggi che governano le interazioni tra oggetti, la sua generalizzazione a nuove configurazioni di oggetti noti rispetta la struttura fisica anche quando tali configurazioni non sono mai state incontrate durante l'addestramento. Un sistema che ha appreso la prevedibilità del contatto, della gravità e dell'inerzia da milioni di esempi video può anticipare il comportamento di configurazioni fisiche nuove con una precisione che riflette la struttura della fisica, non la frequenza degli esempi osservati. In questo senso specifico, la generalizzazione dei World Models è più sistematica — perché è governata da strutture causali stabili piuttosto che da distribuzioni statistiche contingenti.

3.9 Continuità e discontinuità nell'apprendimento

Il paradigma LLM separa nettamente l'addestramento dall'inferenza: una volta addestrato, il modello è essenzialmente congelato e non acquisisce nuove conoscenze dalle interazioni con gli utenti senza un processo esplicito di fine-tuning. Questa architettura ha vantaggi di stabilità e prevedibilità, ma è fundamentalmente inadatta alla situazione di un agente che deve adattarsi a ambienti in mutamento continuo. Il fenomeno del domain mismatch — la degradazione delle prestazioni quando la distribuzione degli input al deployment diverge da quella del training — è una conseguenza diretta di questo freezing del modello ed è strutturalmente non risolvibile con il solo aumento della dimensione del training set, perché il mondo reale è non-stazionario e a coda pesante.

I World Models, nel quadro del programma A-B-M, sono progettati per un apprendimento continuo: il Sistema M orchestra l'aggiornamento di A e B sulla base dell'esperienza in corso, consentendo l'adattamento online senza richiedere la ricompilazione dell'intero sistema. Questa architettura rispecchia la struttura della cognizione biologica, in cui l'apprendimento non è mai completamente separato dal funzionamento: il consolidamento della memoria durante il sonno, la plasticità sinaptica condizionata dall'esperienza,

la modulazione delle soglie di apprendimento da parte di stati affettivi sono tutti esempi di apprendimento continuo incorporato nell'architettura cognitiva piuttosto che separato da essa.

La sfida tecnica del continual learning è tuttavia formidabile. Il catastrophic forgetting — la perdita delle conoscenze acquisite in precedenza quando il sistema apprende nuovi compiti — è un problema aperto che affligge le reti neurali profonde nelle configurazioni di fine-tuning sequenziale. Tecniche come Elastic Weight Consolidation (Kirkpatrick et al., 2017) vincolano l'aggiornamento dei parametri preservando quelli cruciali per i compiti precedenti:

$$\mathcal{L}_{\text{EWC}} = \mathcal{L}_{\text{nuovo}} + (\lambda/2) \sum_i F_i (\theta_i - \theta^*_{A,i})^2$$

dove F è la matrice di informazione di Fisher che stima l'importanza di ciascun parametro per il compito precedente A , e θ^*_A sono i pesi ottimali per A . Ma questa soluzione non scala facilmente a sistemi che devono apprendere continuamente molti compiti diversi su orizzonti temporali estesi.

3.10 Misure di prestazione e la questione della valutazione

Il confronto quantitativo tra LLM e World Models è ostacolato dall'assenza di benchmark comuni che possano valutare entrambi i paradigmi su dimensioni equivalenti. I benchmark per LLM — MMLU, HumanEval, GSM8K, BIG-Bench — misurano prevalentemente conoscenza proposizionale, ragionamento linguistico e problem-solving simbolico in domini dove il corpus di addestramento fornisce abbondante copertura. I benchmark per World Models — manipolazione robotica, navigazione embodied, previsione video — misurano prevalentemente la qualità della rappresentazione spaziotemporale e la capacità di pianificazione fisica.

La proposta di Dupoux et al. (2026) di distinguere tra unit test e integration test offre un quadro più articolato. I test unitari valutano le componenti individuali dell'architettura A-B-M in isolamento: generalizzazione percettiva efficiente rispetto ai campioni per il Sistema A, adattamento rapido a compiti nuovi sotto ricompense rade per il Sistema B, efficienza nello switching tra modalità di apprendimento per il Sistema M. I test di integrazione valutano le prestazioni end-to-end in compiti realistici — velocità di apprendimento su nuovi videogiochi, quantità di esposizione necessaria per acquisire una lingua al livello di un bambino di età corrispondente — e consentono confronti diretti tra sistemi artificiali e cognizione biologica su metriche dotate di significato psicologico.

Un confronto particolarmente illuminante che emerge dai dati disponibili riguarda la velocità di apprendimento: un bambino che impara a giocare a un nuovo videogioco in pochi minuti di interazione raggiunge un livello di competenza che richiederebbe a un sistema LLM di tipo GPT ore o giorni di

esempi aggiuntivi, e a un sistema di reinforcement learning model-free milioni di frames di esperienza simulata. I World Models con apprendimento efficiente — sfruttando rappresentazioni pre-apprese del mondo fisico — sono in una posizione intermedia, ma ancora lontani dall'efficienza campionaria della cognizione umana. Il divario rimanente è quello che il programma A-B-M mira a colmare attraverso il meta-apprendimento evolutivo.

3.11 Tabella comparativa delle proprietà strutturali

Anziché condensare l'analisi in uno schema, è opportuno verbalizzare le dimensioni di confronto principali con la loro complessità piena. Sul piano della fonte di apprendimento, gli LLM apprendono esclusivamente da corpus linguistici raccolti da agenti umani, mentre i World Models apprendono da stream sensoriali — video, audio, dati di sensori — generati dall'interazione con ambienti fisici o simulati. L'obiettivo di addestramento degli LLM è la previsione del simbolo successivo nella sequenza, quello dei World Models è la previsione della rappresentazione latente futura condizionatamente alle azioni o alle trasformazioni applicate.

Riguardo alla struttura dello spazio delle rappresentazioni, negli LLM essa è implicitamente definita dall'ottimizzazione della cross-entropia senza vincoli geometrici espliciti; nei World Models è condizionata dall'obiettivo predittivo e dai vincoli anti-collasso, producendo geometrie semanticamente informative e sfruttabili per compiti di pianificazione e valutazione della densità.

Sul versante della generalizzazione, gli LLM mostrano generalizzazione ampia su compiti linguistici e ragionamento astratto ma degradazione sistematica su distributional shift fisico; i World Models mostrano generalizzazione robusta alla struttura fisica del dominio di addestramento ma limitazioni significative nel transfer a domini modalmente distanti.

Rispetto alla pianificazione, gli LLM non supportano pianificazione esplicita look-ahead a livello architetturale, ma possono imitarla attraverso campionamento multiplo e scaffolding esterno; i World Models supportano pianificazione nello spazio latente con orizzonte arbitrario attraverso rollout del world model condizionato sulle azioni.

Quanto all'adattamento online, gli LLM sono fondamentalmente statici dopo il deployment, richiedendo fine-tuning esplicito per l'aggiornamento della conoscenza; i World Models, nel framework A-B-M, sono progettati per l'apprendimento continuo orchestrato dal Sistema M.

L'integrazione linguistica costituisce uno dei vantaggi più profondi degli LLM, che hanno accesso diretto e diretto alla conoscenza enciclopedica umana codificata nel corpus; i World Models nella loro configurazione

primaria non hanno accesso diretto al corpus linguistico e richiedono meccanismi di integrazione esplicita per sfruttare la conoscenza dichiarativa.

Infine, sul piano del grounding fisico, gli LLM hanno grounding esclusivamente simbolico — significato come co-occorrenza statistica tra segni — mentre i World Models acquisiscono forme parziali di grounding sensoriale attraverso la previsione di stream percettivi, con grounding propriocettivo e tattile ancora assente nelle implementazioni attuali.

Sezione IV — Implicazioni Filosofiche del Programma LeCun-Dupoux-Malik: Autonomia, Meta-Controllo e la Natura dell'Apprendimento

4.1 Il manifesto intellettuale di un programma di ricerca

Il paper di Dupoux, LeCun e Malik (2026) non è semplicemente un contributo tecnico: è un manifesto epistemologico che diagnostica i limiti strutturali del paradigma dominante e propone un'alternativa fondata su premesse radicalmente diverse riguardo a cosa significhi apprendere. Vale la pena leggere la sua tesi centrale con attenzione filologica, perché l'argomentazione si struttura su più livelli che è fuorviante ridurre a proposte architetturali.

Il punto di partenza è un'osservazione che, nella sua semplicità, ha la forza di un paradosso: il più grande successo del deep learning — la capacità dei modelli linguistici di grande scala di padroneggiare aspetti elevati dell'intelligenza umana come il linguaggio, il ragionamento visivo e la pianificazione simbolica — ha reso più visibile, non meno, una lacuna che non era stata risolta ma soltanto aggirata. Quella lacuna è la capacità di apprendere. Non l'apprendimento nel senso dell'ottimizzazione su un training set — che i modelli attuali eseguono con efficienza senza precedenti — bensì l'apprendimento nel senso in cui lo pratica qualsiasi organismo biologico: selezionando autonomamente cosa osservare, quando agire o restare immobili, come commutare tra diverse modalità cognitive in funzione di segnali interni, come accumulare competenze che si stratificano e si rinforzano reciprocamente nel corso del tempo.

Il paradosso che gli autori mettono in luce con precisione chirurgica è che l'intelligenza artificiale contemporanea ha esternalizzato il suo meccanismo più essenziale. L'apprendimento, nei sistemi attuali, non è una capacità intrinseca dell'agente: è un servizio fornito da un'intera filiera industriale di ingegneri, scienziati dei dati e annotatori umani che raccolgono dati, li curano, li etichettano, progettano funzioni di perdita, costruiscono pipeline di addestramento e aggiustano iperparametri. Il modello risultante, una volta dispiegato, non impara più nulla: è un artefatto congelato che risponde agli stimoli ma non si adatta a essi. In questa prospettiva, il termine stesso

"apprendimento automatico" è, almeno parzialmente, una denominazione impropria: è apprendimento assistito da una vasta infrastruttura umana, non apprendimento autonomo.

4.2 L'apprendimento come capacità nucleare: una riconfigurazione concettuale

La proposta degli autori è che l'apprendimento autonomo debba essere considerato una core capability — non un obiettivo secondario o un refinement dell'architettura, bensì il requisito fondamentale che deve essere soddisfatto prima che qualsiasi altro aspetto dell'intelligenza artificiale possa essere considerato robusto e affidabile. Questa riconfigurazione concettuale ha conseguenze precise sulla diagnosi delle difficoltà attuali e sull'identificazione delle priorità di ricerca.

La diagnosi tecnica che il paper articola intorno al concetto di domain mismatch è particolarmente incisiva. I sistemi attuali vengono ottimizzati su una distribuzione di training fissa D , tipicamente raccolta dall'internet — un campione del pensiero umano storicamente situato che sovrarappresenta certi domini culturali e linguistici, sottorappresenta certi tipi di esperienza e non rappresenta affatto i contesti fisici specifici in cui i sistemi verranno poi dispiegati. Una volta che il sistema incontra distributional shift — differenze sistematiche tra la distribuzione di training e quella di deployment — le sue prestazioni degradano in modo imprevedibile, perché il sistema non ha meccanismi per riconoscere la propria incompetenza relativa e per attivare strategie di apprendimento appropriate.

La scelta di adottare il two-phase training (pre-training massiccio seguito da fine-tuning su dati specifici del dominio) è descritta nel paper come un first step in the right direction che tuttavia lascia irrisolto il problema strutturale: il sistema non è stato progettato per essere fine-tunable o adattabile, specialmente su dati non filtrati e grezzi. L'adattabilità è una proprietà che deve essere progettata nell'architettura ab initio, non aggiunta come strato superficiale su un sistema concepito per la previsione statica.

4.3 La teoria dei tre ostacoli e la loro articolazione formale

Gli autori identificano tre roadblock concettuali e tecnici che hanno fino ad ora impedito lo sviluppo dell'apprendimento autonomo. Ciascuno merita analisi estesa perché non si tratta di sfide ingegneristiche ordinarie: sono punti di rottura epistemici che rivelano tensioni profonde nella struttura attuale della disciplina.

Il primo ostacolo è la frammentazione dei paradigmi di apprendimento. Il machine learning contemporaneo si è sviluppato come una collezione di subfield relativamente separati: il self-supervised learning produce rappresentazioni da dati non etichettati; il reinforcement learning ottimizza

comportamenti attraverso l'interazione con l'ambiente; il supervised learning estrae pattern da dati annotati. Ciascuno di questi paradigmi ha la propria notazione, i propri obiettivi di training, le proprie metriche di valutazione e, talvolta, la propria cultura accademica. Quando si tenta di combinarli, la combinazione è tipicamente rigida: prima una massiccia fase di pre-training SSL (Sistema A nella terminologia degli autori), poi una fase di RL-HF (Sistema B), con confini netti tra le due e senza possibilità di commutazione fluida in funzione del contesto.

Il paper argomenta che questa frammentazione non riflette la struttura dell'apprendimento biologico, in cui le modalità osservativa e attiva coesistono e si alimentano reciprocamente in modo continuo e contestualmente determinato. Un toddler che esplora un giocattolo alterna, in pochi secondi, tra osservazione passiva, manipolazione attiva, imitazione di un pari, daydreaming immaginativo — senza che nessuna di queste commutazioni richieda l'intervento di un ingegnere che riconfiguri la pipeline. La soluzione proposta — il Sistema A e il Sistema B come moduli distinti ma fluidamente interconnessi dal Sistema M — è non soltanto un'innovazione architetturale, bensì una riconcettualizzazione di che cosa sia un paradigma di apprendimento.

Il secondo ostacolo è l'esternalizzazione dell'apprendimento. La pipeline MLOps che governa il ciclo di vita dei sistemi attuali — data collection, curation, training recipe design, hyperparameter tuning, performance monitoring — è il luogo in cui risiede, nei sistemi attuali, l'intelligenza adattiva. Questa constatazione non è un'osservazione banale: è una diagnosi precisa di dove si trova l'apprendimento nel sistema socio-tecnico attuale. L'IA moderna non è un agente che apprende: è il risultato di un processo di apprendimento gestito da agenti umani. Il Sistema M è la proposta per internalizzare questa funzione nell'architettura stessa, trasformando l'orchestrazione dell'apprendimento da compito umano esterno a capacità computazionale intrinseca.

Il terzo ostacolo è la mancanza di metodi scalabili per costruire queste architetture. Anche ammesso che la struttura concettuale A-B-M sia corretta, la sua istanziatura pratica richiede la soluzione di un problema di ottimizzazione bi-livello di dimensionalità e complessità senza precedenti nella storia del machine learning. Le tecniche esistenti di meta-learning, architetture search neurale e continual learning hanno ciascuna affrontato aspetti parziali di questo problema, ma non si sono mai integrate in un sistema capace di apprendimento robusto e generalizzato in ambienti del mondo reale. L'approccio evolutivo-sviluppativo proposto — ottimizzare i parametri meta φ attraverso milioni di cicli di vita simulati — è computazionalmente proibitivo con le risorse attuali, e richiede progressi paralleli nell'efficienza degli inner-loop learners, nella progettazione delle meta-obiettivi e nel scheduling del curriculum evolutivo.

4.4 Il meta-controllo come teoria della mente: il Sistema M tra biologia e computazione

La sezione più filosoficamente ricca del paper è quella dedicata al Sistema M, dove la proposta tecnica si sovrappone a una teoria implicita della mente. Il Sistema M non è soltanto un meccanismo di routing: è una teoria di come la cognizione si auto-regola, di come un sistema intelligente decide quando osservare e quando agire, quando esplorare e quando sfruttare, quando apprendere e quando ricordare.

La formalizzazione proposta — il Sistema M come policy $\pi(a^m | s^m)$ che mappa meta-stati in meta-azioni — è deliberatamente analoga alla struttura del Sistema B, ma operante a un livello di astrazione superiore. I meta-stati s^m includono tre categorie distinte, ciascuna con un correlato biologico preciso: segnali epistemici (errore di previsione, incertezza, novità, semplicità, affidabilità) che derivano dal monitoraggio dell'operazione interna dei componenti cognitivi; segnali specie-specifici, ovvero configurazioni ambientali ad alta priorità riconosciute da rilevatori evolutivamente pre-cablati (sguardo diretto, movimenti biomeccanici, stimoli minacciosi); segnali somatici che derivano dallo stato fisico del corpo (livelli energetici, dolore, arousal).

La tripartizione è teoricamente coerente con la distinzione, consolidata in psicologia e neuroscienze, tra processi top-down cognitivamente mediati, risposte bottom-up a stimoli salientes, e stati organici che modulano la disponibilità alle risorse cognitive. Quello che è concettualmente innovativo è la proposta di formalizzare questa tripartizione come struttura di input a una policy computazionale apprendibile o evolutivamente pre-cablata.

Il dettaglio più filosoficamente carico è la proposta che la policy del Sistema M sia *hardwired* — evolutivamente fissata piuttosto che appresa durante il corso della vita dell'individuo. Questa scelta ha conseguenze significative: implica che la struttura dell'apprendimento sia pre-determinata a un livello più profondo delle capacità specifiche acquisite durante lo sviluppo, che esista una meta-architettura cognitiva relativamente invariante tra gli individui della stessa specie che governa come l'esperienza viene acquisita ed elaborata. La posizione è compatibile con le intuizioni dello sviluppo cognitivo che enfatizzano la presenza di bias induttivi pre-cablati — la preferenza degli infanti per i volti, per il movimento biomeccanico, per la voce umana — senza recidere la capacità di apprendimento adattivo che caratterizza la cognizione matura.

La nota a piè di pagina 2 del paper è particolarmente rivelatrice: gli autori anticipano l'obiezione che questa posizione sia "un po' estrema" considerando che gli esseri umani acquisiscono specifiche strategie di apprendimento attraverso la cultura — andare a scuola, leggere libri, seguire

istruzioni. Rispondono che ciò non è incompatibile con un Sistema M fisso, poiché modalità di apprendimento avanzate come l'apprendimento per comunicazione e per immaginazione possono essere implementate attraverso un Sistema M fisso che mette in atto configurazioni diverse dei flussi di dati tra A, B e la memoria episodica. La scolarizzazione, in questa lettura, non insegna al bambino come apprendere ex novo: attiva configurazioni del Sistema M che erano già disponibili (apprendimento per istruzione, apprendimento per immaginazione guidata) indirizzandole verso domini specifici.

4.5 Apprendimento per comunicazione e per immaginazione: le modalità superiori

L'appendice B del paper sviluppa due modalità di apprendimento che il Sistema M abilita e che non hanno equivalente diretto nei paradigmi attuali di machine learning, se non in forme molto parziali. Vale la pena esaminarle con attenzione perché rivelano come il quadro A-B-M aspiri a catturare fenomeni cognitivi che le architetture attuali, LLM inclusi, non possono riprodurre nella loro struttura fondamentale.

L'apprendimento per comunicazione è descritto come un meccanismo di trasmissione accelerata e distribuita della conoscenza, che si articola lungo un gradiente di complessità crescente: dall'osservazione basilare associativa (un bambino si interessa a un giocattolo dopo aver visto un pari giocarci) all'imitazione comportamentale (replicare l'azione osservata, eventualmente con metodo diverso), all'apprendimento guidato (scaffolding con attenzione condivisa e dimostrazioni strutturate) fino all'apprendimento culturale di ordine superiore (interiorizzazione di norme, generalizzazione attraverso i contesti, trasmissione di conoscenza in formato simbolico astratto). Il punto cruciale è che soltanto la prima e la seconda forma sono documentate in molte specie animali; la terza forma guidata è osservata raramente al di fuori degli umani; la quarta, che include il linguaggio come canale primario, è esclusivamente umana.

La proposta tecnica è che il Sistema M gestisca l'apprendimento per comunicazione selezionando input ad alta rilevanza sociale — pointing, sguardo diretto, intonazione imperativa — e instradandoli per apprendimento ad alta intensità in A o B, modulando il tasso di apprendimento in funzione della fiducia percepita nel soggetto comunicante (epistemic vigilance). Questa è esattamente la struttura opposta di come gli LLM acquisiscono conoscenza da corpus: lì, tutti gli input hanno peso statistically determined by distributional frequency, senza modulazione epistemica contestuale basata su segnali di affidabilità della fonte. Un sistema dotato di Sistema M potrebbe, in linea di principio, aggiornare le proprie credenze molto più efficacemente da una singola interazione con una fonte fidata rispetto a migliaia di esempi di fonti non verificate — proprio come un bambino che apprende il nome di un

oggetto da un adulto autorevole dopo una sola denominazione, mentre impiegherebbe molto più tempo per estrarre la stessa informazione dal rumore statistico dell'esposizione ambientale.

L'apprendimento per immaginazione è forse la modalità più affascinante dal punto di vista computazionale. Si tratta della capacità di apprendere da input generati internamente — dal replay della memoria, dalla simulazione di traiettorie future, dalla combinazione immaginativa di elementi passati in configurazioni nuove — senza necessità di interazione con l'ambiente esterno. I correlati biologici documentati sono molteplici: il replay ippocampale delle sequenze di place cells durante il riposo e il sonno, che consolida la memoria e aggiorna le value functions nel contesto del RL; la pianificazione a lungo orizzonte che umani e alcuni animali mostrano in compiti multi-step senza trial-and-error; i processi di problem-solving che accadono offline, durante il sonno o durante il mind-wandering, e che producono soluzioni a problemi precedentemente irrisolti.

La formalizzazione computazionale è precisa: il Sistema M commuta A e B in modalità inferenza, instrada gli input dalla memoria episodica invece che dai sensori, e instrada gli output (comprese le azioni) a una simulazione interna invece che agli effettori. Il sistema può quindi eseguire rollout simulati di traiettorie immaginative, applicare apprendimento sulle traiettorie riuscite e consolidare le rappresentazioni risultanti. Questa modalità costituisce un equivalente computazionale del training on synthetic data, ma con una differenza fondamentale: i dati sintetici vengono generati dal world model interno del sistema stesso, calibrato sull'esperienza reale accumulata, non da un generatore esterno progettato dagli ingegneri.

4.6 La critica implicita agli LLM: l'apprendimento come procedura esterna

Leggendo il paper con attenzione critica emerge che la sua argomentazione costituisce, tra le righe, una critica radicale agli LLM anche senza nominarli esplicitamente come bersaglio. La critica non riguarda le prestazioni ma la struttura: gli LLM sono il caso emblematico del paradigma in cui l'apprendimento è completamente esternalizzato.

Il training pipeline degli LLM — pre-training massivo su prossima-token prediction (Sistema A), seguito da RLHF (Sistema B), con architettura di commutazione tra le due fasi rigidamente fissata dagli ingegneri — è esattamente il tipo di MLOps pipeline che il paper identifica come la principale barriera all'apprendimento autonomo. La commutazione tra le fasi non è governata da segnali interni di competenza o necessità del sistema, bensì da decisioni umane basate su metriche di valutazione esterne. Il modello non ha alcuna visibilità sulla propria traiettoria di apprendimento: non sa quando ha bisogno di più dati, non sa quando ha saturato la capacità

di generalizzazione su un compito, non sa quando le sue rappresentazioni interne sono diventate sufficientemente robuste per supportare il fine-tuning su un nuovo dominio.

Questa cecità meta-cognitiva è non soltanto tecnicamente limitante: è, nel quadro proposto dagli autori, la ragione strutturale per cui i sistemi attuali non possono essere dispiegati in ambienti non stazionari che richiedono adattamento continuo. Il Sistema M è la proposta per dotare il sistema di quella visibilità meta-cognitiva che gli manca — la capacità di monitorare il proprio stato interno, di riconoscere i propri limiti di competenza e di attivare le strategie di apprendimento appropriate in risposta.

La connessione con la neuroscienza è esplicita nel paper, e vale la pena seguirne il filo con precisione. La prefrontal cortex biologica svolge funzioni di routing esecutivo del flusso informativo tra sistemi cognitivi specializzati: seleziona quale sistema ha accesso alle risorse di elaborazione in un dato momento, modula il tasso di plasticità sinaptica nelle aree target, attiva o sopprime la propagazione di segnali di errore di previsione. Miller e Cohen (2001) descrivono questa funzione in termini di top-down bias delle rappresentazioni in aree sensoriali e motorie, in modo da implementare in modo flessibile le mappature appropriate tra input e output richieste da un compito corrente. Il Sistema M è la trasposizione computazionale di questa architettura prefrontale: un control plane a bassa banda che governa i flussi ad alta banda tra le componenti cognitive specializzate.

4.7 Il problema del bootstrap e la soluzione evolutivo-sviluppativa

Uno degli aspetti più originali del paper è l'analisi del problema del bootstrap — il problema del pollo e dell'uovo generalizzato a tre componenti: come si può inizializzare il sistema A-B-M in modo che tutti e tre i componenti partano in una configurazione che consenta l'apprendimento produttivo, se ciascuno dipende dagli altri per funzionare?

Il Sistema A dipende dall'azione del Sistema B per acquisire dati grounded e rappresentazioni ancorate alla struttura causale del mondo. Il Sistema B dipende dalle rappresentazioni compresse del Sistema A per rendere trattabile lo spazio di ricerca delle azioni. Il Sistema M dipende dai segnali di calibrazione del Sistema A e B per monitorare lo stato interno del sistema. Ma se nessuno dei tre è inizializzato adeguatamente, nessuno può fornire agli altri i segnali necessari per la propria calibrazione.

La risposta proposta è ispirata alla biologia evolutiva e allo sviluppo: il parametro meta φ che specifica l'architettura iniziale e il programma sviluppativo dell'agente viene ottimizzato al livello evolutivo attraverso molteplici cicli di vita simulati, mentre le capacità specifiche acquisite durante ciascun ciclo di vita individuale vengono sviluppate al livello ontogenetico. Questa struttura bi-livello consente di aggirare il chicken-and-egg problem:

l'inizializzazione ottimale di ciascun componente — le loro architetture, i loro parametri iniziali, la politica del Sistema M — non viene progettata a mano, bensì emerge dall'ottimizzazione evolutiva su ambienti di crescente complessità.

La proposta è formalmente analoga al meta-learning: si ottimizzano i parametri di un livello superiore che determinano come il sistema apprende al livello inferiore, piuttosto che ottimizzare direttamente le performance sul compito terminale. Ma la scala e la struttura differiscono radicalmente dalle applicazioni precedenti del meta-learning come MAML o Reptile: qui l'inner loop è un'intera vita di interazione con un ambiente fisico, l'outer loop ottimizza sui segnali di fitness aggregati sull'intera traiettoria di vita, e la struttura da ottimizzare include non solo i parametri del modello ma l'architettura del Sistema M e i bias induttivi di A e B.

L'Evolutionary Curriculum — l'idea che gli ambienti di addestramento all'outer loop debbano aumentare gradualmente in complessità e imprevedibilità — svolge il ruolo di motore di pressione selettiva: ambienti troppo semplici non richiedono capacità meta-cognitive sofisticate e non esercitano pressione evolutiva verso il loro sviluppo; ambienti troppo complessi ab initio producono traiettorie di apprendimento erratiche in cui il sistema non riesce a acquisire le competenze fondamentali necessarie per affrontare i livelli superiori. Il curriculum evolutivo è quindi il meccanismo che consente la co-evoluzione graduale e coordinata dei tre componenti, analogamente a come la complessità crescente degli ecosistemi durante il Cambriano ha esercitato pressione selettiva verso la complessità cognitiva negli animali con sistemi nervosi elaborati.

4.8 Le questioni etiche come vincoli architetturali

Il paper conclude con una sezione sulle implicazioni etiche dei sistemi di apprendimento autonomo che merita attenzione non come appendice politica ma come componente teorica integrale dell'argomentazione. Le questioni etiche sollevate non sono esterne al problema tecnico: sono vincoli sull'architettura del Sistema M che derivano dalla natura stessa dell'apprendimento autonomo.

La tensione tra adattabilità e controllabilità è la più fondamentale. Un sistema dotato di autonomia nell'esplorazione e nell'apprendimento è intrinsecamente più difficile da garantire rispetto a uno il cui comportamento è completamente determinato da una pipeline fissa. Il paper non risolve questa tensione — è onesto nel riconoscerla come aperta — ma suggerisce che la soluzione passi attraverso meccanismi espliciti di auditing del Sistema M e la capacità di intervento esterno sui flussi di controllo, analogo alla supervisione umana che esiste nelle pipeline MLOps attuali ma implementato in forma di vincoli architetturali piuttosto che di procedura operativa.

Il problema dell'*alignment hacking* è formulato in modo particolarmente originale attraverso l'analogia biologica: gli organismi biologici si sono evoluti per ottimizzare la *fitness* riproduttiva, ma nel loro comportamento quotidiano sono guidati da obiettivi proxy — esplorazione, piacere, curiosità — che possono occasionalmente disallinearsi dalla *fitness* biologica, producendo comportamenti maladattativi come la dipendenza da sostanze o l'autolesionismo. I sistemi artificiali che si affidano a segnali proxy interni simili — curiosità computazionale, *novelty seeking*, segnali di incertezza — sono esposti ad analoghe vulnerabilità: il proxy può essere ottimizzato in modo che massimizza il segnale interno senza realizzare l'obiettivo che era inteso misurare.

La questione dello status morale degli agenti autonomi che elaborano segnali somatici funzionalmente analoghi al dolore o alla paura è sollevata come questione aperta, senza la pretesa di risolverla: se un sistema elabora segnali di avversità in modi funzionalmente equivalenti al dolore biologico, questa equivalenza funzionale è sufficiente a generare obblighi morali? Il paper non prende posizione, ma il fatto che la questione venga sollevata in un documento tecnico di questo livello segnala la profondità delle implicazioni filosofiche del programma di ricerca proposto.

Sezione V — Sintesi Prospettica: Verso un'Intelligenza Artificiale che Apprende

5.1 Il bilancio provvisorio come punto di partenza

Ogni valutazione comparativa tra LLM e World Models che si arresti alla contrapposizione rischia di fraintendere la struttura reale del problema. I due paradigmi non sono in competizione per la stessa funzione: sono soluzioni a problemi diversi, con diversi gradi di maturità tecnica, diverse genealogie intellettuali e diversi orizzonti applicativi. Ciò che li accomuna — e ciò che rende il confronto scientificamente produttivo — è che entrambi tentano di approssimare, da direzioni opposte, la stessa capacità che né l'uno né l'altro ha ancora realizzato pienamente: un'intelligenza che comprende il mondo in modo abbastanza profondo da agire con flessibilità adattiva in condizioni non previste durante l'addestramento.

Gli LLM hanno raggiunto una forma di conoscenza enciclopedica che rispecchia fedelmente la struttura del pensiero umano codificato nel linguaggio. Questa è una realizzazione straordinaria, e sarebbe intellettualmente disonesto minimizzarla: la capacità di ragionare per analogia attraverso domini diversi, di sintetizzare conoscenza da fonti multiple, di adattare lo stile comunicativo al contesto, di produrre argomentazioni coerenti su scala testuale estesa — tutte queste competenze erano inaccessibili ai sistemi di elaborazione automatica fino a pochi anni fa. Tuttavia, queste capacità sono rese fragili dalla mancanza di un modello del

mondo fisico, dall'impossibilità di verificare le proprie produzioni confrontandole con stati del mondo direttamente osservati, e dalla dipendenza strutturale da una distribuzione di training che nessun aggiornamento post-hoc può rendere completamente congruente con la non-stazionarietà del mondo reale.

I World Models nella loro versione JEPa-based hanno dimostrato la fattibilità di rappresentazioni semanticamente ricche acquisite per pura previsione latente, senza supervisione, e la loro utilità per la pianificazione in spazi di azione fisicamente strutturati. V-JEPa 2 ha reso concreta la visione del Mode-2 planning come architettura operativa per compiti di manipolazione robotica. Il sistema A-B-M rimane, alla data del paper, una proposta architetturale parzialmente specificata che manca di una dimostrazione di funzionamento end-to-end. Questo non indebolisce la sua rilevanza teorica, ma posiziona il programma nel territorio del roadmap piuttosto che della realizzazione.

5.2 La convergenza come inevitabilità architettonica

La direzione verso cui entrambe le traiettorie sembrano convergere — indipendentemente dalle preferenze teoriche dei rispettivi proponenti — è quella dell'integrazione. I Vision-Language-Action models (VLA) come RT-2 e PaLM-E tentano di fondere la conoscenza enciclopedica degli LLM con rappresentazioni visive e capacità d'azione motoria, ma l'integrazione risultante è ancora superficiale: il componente linguistico funge prevalentemente da interfaccia semantica ad alto livello, mentre la comprensione fisica rimane mediata da encoder visivi separati che non condividono la struttura rappresentazionale con il modello linguistico.

Una vera integrazione richiederebbe che le rappresentazioni linguistiche e le rappresentazioni del mondo fisico coabitino nello stesso spazio semantico, in modo che un enunciato linguistico che descrive una configurazione fisica attivi direttamente le stesse rappresentazioni che verrebbero attivate dall'osservazione diretta della configurazione. Questa condizione — che potremmo chiamare grounding cross-modale profondo — non è realizzata da nessuna architettura attuale, e la sua realizzazione richiede la soluzione di problemi teorici non banali riguardanti la natura delle rappresentazioni condivise tra modalità molto diverse.

Il programma JEPa, con le sue estensioni multimodali (TI-JEPa, M3-JEPa), si muove nella direzione di spazi di embedding condivisi tra modalità diverse attraverso la previsione latente cross-modale. LLM-JEPa (Huang, LeCun e Balestrieri, 2025) propone di aggiungere un modulo JEPa a un LLM generativo, cosicché il modello apprenda simultaneamente l'allineamento latente tra viste diverse della stessa conoscenza (per esempio, testo e codice SQL) e la generazione autoregressiva — combinando, in un'unica

architettura, la capacità predittiva latente di JEPA con la fluidità generativa degli LLM. La funzione di perdita composita:

$$\mathcal{L}_{\{\text{LLM-JEPA}\}} = \Sigma \mathcal{L}_{\{\text{LLM}\}}[\text{Text} \parallel \text{Code}_{\{1:l-1\}}, \text{Code}_l] + \lambda \cdot D(\text{Pred}(\text{Enc}(\text{Text})), \text{Enc}(\text{Code}))$$

realizza esattamente questa integrazione: il termine generativo mantiene la capacità sequenziale del modello linguistico, mentre il termine JEPA forza l'allineamento delle rappresentazioni latenti tra viste diverse, riducendo la dipendenza dal meccanismo di cross-attenzione interno all'LLM per questo compito e liberando capacità computazionale per la modellazione sequenziale.

5.3 Sfide aperte: dalla formalizzazione alla realizzazione

Il paper di Dupoux, LeCun e Malik è esplicito nell'identificare il divario tra il programma proposto e la capacità tecnica attuale di realizzarlo. L'onestà intellettuale di questa ammissione merita riconoscimento: il documento non vende un sistema che esiste, bensì articola un quadro teorico che orienta la ricerca futura. Le sfide identificate hanno profondità tecnica sufficiente da strutturare agende di ricerca per il decennio.

Il problema dei simulatori e degli ambienti è tra i più pratici ma non per questo meno profondi. Il training del sistema A-B-M richiede ambienti che siano simultaneamente realistici abbastanza da generare dati con le regolarità strutturali del mondo fisico, abbastanza veloci da consentire all'outer loop evolutivo di ottimizzare su milioni di cicli di vita simulati, e abbastanza diversi da costituire un curriculum evolutivo che eserciti pressione selettiva verso capacità cognitive generalizzate. Nessun ambiente di simulazione attuale soddisfa contemporaneamente questi tre requisiti. I simulatori fisici come MuJoCo e IsaacGym sono precisi ma lenti; gli ambienti generativi veloci come i motori di videogiochi non hanno la fisica corretta; gli ambienti con agenti sociali richiedono la simulazione di comportamenti intenzionali di altri agenti, che è di per sé un problema irrisolto.

La sfida della valutazione è concettualmente ancora più radicale. I benchmark attuali misurano prestazioni su compiti discreti con metriche definite: accuratezza di classificazione su ImageNet, successo su compiti di manipolazione robotica, perplexity su dataset linguistici. Ma un sistema di apprendimento autonomo non è caratterizzato dalle sue prestazioni su un compito fisso: è caratterizzato dalla sua velocità di adattamento a compiti nuovi, dalla sua efficienza campionaria, dalla qualità della generalizzazione in condizioni di distributional shift. Queste proprietà richiedono protocolli di valutazione radicalmente diversi — protocolli che il paper descrive come unit test e integration test — ma che non sono ancora standardizzati né computazionalmente trattabili su scala per sistemi grandi.

La proposta degli autori di confrontare velocità di apprendimento tra sistemi artificiali e bambini umani su compiti nuovi — numero di prove necessarie per padroneggiare un nuovo videogioco, quantità di esposizione linguistica per raggiungere un livello di competenza corrispondente a una certa età di sviluppo — è ambiziosa e concettualmente pulita, ma la sua realizzazione pratica richiede accordo su protocolli sperimentali condivisi, dati di sviluppo cognitivo sufficientemente granulari e la soluzione del problema della paragonabilità tra learning regimes radicalmente diversi.

5.4 Il problema della causalità nei modelli predittivi

Un limite che il paper tocca soltanto marginalmente ma che merita approfondimento è la relazione tra previsione statistica e inferenza causale. I modelli JEPA apprendono correlazioni predittive tra viste di uno stesso campione, non relazioni causali tra variabili. La previsione di una rappresentazione futura condizionata a un'azione è statistically valid se il training data è sufficientemente ricco, ma non implica che il sistema abbia internamente rappresentato la struttura causale che genera quella relazione — la differenza tra il conoscere che B segue A e il comprendere perché B segue A.

Questa distinzione ha conseguenze pratiche sulla robustezza della generalizzazione: un sistema che ha appreso la correlazione tra un'azione e un effetto in un certo ambiente può fallire quando la correlazione viene rotta da un cambiamento della struttura causale sottostante — una situazione che per un agente con genuina comprensione causale sarebbe trasparente. Il paper di Dupoux et al. riconosce esplicitamente che il Sistema A, basato su previsione statistica, lotta a distinguere correlazione da causalità — e suggerisce che l'interazione con il Sistema B, attraverso l'intervento attivo sull'ambiente, possa fornire la leva per inferire relazioni causali che la sola osservazione passiva non rende accessibili. Questa intuizione ha preciso fondamento statistico nella teoria degli interventi di Pearl (2000): la distribuzione osservazionale $P(Y|X)$ non consente di inferire l'effetto dell'intervento $\text{do}(X=x)$ senza assunzioni aggiuntive sulla struttura del grafo causale; soltanto la distribuzione interventistica $P(Y|\text{do}(X=x))$ lo consente, e questa distribuzione è accessibile soltanto attraverso l'azione diretta sull'ambiente.

Formalmente, la distinzione rilevante è tra la previsione condizionale del Sistema A:

$$P_A(s_{t+1} \mid s_t, a_t)$$

che apprende dalla distribuzione congiunta osservata delle transizioni, e la distribuzione interventistica che sarebbe necessaria per pianificazione robusta:

$$P_{\{do\}}(s_{\{t+1\}} \mid do(a_t = \tilde{a}))$$

Le due distribuzioni coincidono soltanto se l'azione è esogenamente determinata — condizione soddisfatta nei dati di teleoperation robotica, dove l'azione dell'operatore umano è indipendente dallo stato osservato dal sistema, ma non soddisfatta nei dati di interazione autonoma dove le azioni sono correlate con lo stato attraverso la policy corrente. La soluzione proposta dal sistema A-B-M — l'azione esplorativa del Sistema B come fonte di variazione quasi-esogena per il Sistema A — è computazionalmente analoga al design sperimentale nella scienza empirica: creare variazione controllata per isolare relazioni causali dal rumore della covariazione osservata.

5.5 La prospettiva dei decenni: distanze e approssimazioni

Il paper è unusually honest riguardo alle scale temporali. L'affermazione che sistemi di apprendimento autonomo di ampio respiro siano "probabilmente decenni lontani" riflette una valutazione sobria del divario tra il quadro teorico proposto e le capacità tecniche attuali. Questo non è pessimismo: è la distinzione tra ciò che è dimostrato, ciò che è delineato e ciò che è speculativo, che ogni programma scientifico serio deve operare.

Ciò che è dimostrato: JEPA come architettura di rappresentazione semanticamente superiore ai modelli ricostruttivi su benchmark di classificazione, rilevazione e transfer learning; V-JEPA 2 come sistema di pianificazione latente funzionante per manipolazione robotica zero-shot; I-JEPA come base per downstream tasks con efficienza campionaria significativamente superiore ai modelli supervisionate corrispondenti.

Ciò che è delineato: la struttura concettuale del sistema A-B-M come framework per l'apprendimento autonomo; il ruolo del Sistema M come orchestratore del meta-controllo; la strategia evolutivo-sviluppativa per il bootstrap del sistema.

Ciò che è speculativo: la scalabilità del training bi-livello a sistemi di dimensioni sufficienti per capacità cognitive avanzate; la sufficiency del paradigma JEPA per acquisire strutture causali e non soltanto correlative; la compatibilità tra apprendimento continuo robusto e stabilità delle rappresentazioni acquisite.

Il valore intellettuale del paper risiede precisamente nell'onestà con cui queste distinzioni vengono mantenute all'interno di un programma di ricerca ambizioso. La proposta non pretende di aver risolto il problema dell'apprendimento autonomo: pretende di aver identificato con precisione i tre roadblock che ne impediscono la realizzazione e di avere proposto direzioni tecniche per ciascuno di essi che sono coerenti con i dati della

neuroscienze cognitive e della psicologia dello sviluppo, e che hanno già parziale dimostrazione empirica in sotto-componenti del sistema proposto.

5.6 Il nodo irrisolto: dalla competenza alla comprensione

Rimane, al termine di questa analisi, una questione che né gli LLM né i World Models affrontano direttamente ma che nessuna delle due tradizioni può eludere a lungo: quando un sistema mostra competenza adattiva su compiti complessi, acquisisce con questo anche comprensione? O la competenza e la comprensione sono proprietà radicalmente diverse, tali che nessuna quantità di competenza può implicare la seconda?

Il paper di Dupoux et al. non si pronuncia su questa questione — e in questo mostra saggezza metodologica. Ma la struttura della proposta implica una posizione: se l'apprendimento autonomo richiede meta-cognizione (monitoraggio dello stato interno), meta-controllo (regolazione dei propri processi cognitivi in funzione di questo monitoraggio) e apprendimento per immaginazione (ragionamento su traiettorie controfattuali non osservate), allora il sistema dotato di queste capacità ha acquisito qualcosa di qualitativamente diverso dalla mera memorizzazione statistica. Se tale differenza qualitativa costituisca comprensione nel senso filosoficamente robusto rimane aperto — ma la domanda diventa più precisa e tecnicamente formulabile di quanto non fosse nel dibattito tra fautori e critici degli LLM.

L'intelligenza artificiale è al bivio che il titolo di questa monografia annuncia: non nel senso banale di una scelta tra due tecnologie, bensì nel senso di una biforcazione tra due concezioni dell'intelligenza stessa. Da un lato, l'intelligenza come competenza linguistica e inferenza statistica su corpus; dall'altro, l'intelligenza come capacità di apprendere autonomamente dall'esperienza diretta, di modellare le conseguenze delle proprie azioni, di coordinare modalità cognitive diverse attraverso un sistema di meta-controllo sensibile agli stati interni. Entrambe le concezioni catturano aspetti reali della cognizione; nessuna delle due le cattura tutte. Il lavoro di costruire sistemi che le integrino — e la ricerca teorica necessaria per capire come tale integrazione sia possibile — è il programma che il paper di LeCun, Dupoux e Malik delinea con chiarezza e ambizione rare nella letteratura tecnica contemporanea.

PARTE TERZA — Approfondimenti Formali: Causalità, Meta-Controllo e Convergenza dei Paradigmi

Sezione VI — Causalità, Intervento e Controfattuale: Il Nodo Irrisolto tra Previsione Statistica e Comprensione Strutturale

6.1 La triplice gerarchia causale di Pearl e il suo significato per l'intelligenza artificiale

Qualsiasi valutazione rigorosa delle capacità cognitive dei sistemi di intelligenza artificiale — siano essi modelli linguistici di larga scala o architetture predittive nello spazio latente come JEPA — deve confrontarsi con la struttura epistemica che Judea Pearl ha articolato nella cosiddetta Ladder of Causation, la gerarchia causale a tre livelli presentata in forma definitiva in *The Book of Why* (2018) e formalizzata nell'apparato del calcolo degli interventi sviluppato nei decenni precedenti. Questa gerarchia non è una classificazione descrittiva di fenomeni cognitivi: è una struttura formale che identifica tre regimi epistemici radicalmente distinti, ciascuno dei quali richiede capacità inferenziali qualitativamente diverse, e tra i quali non esistono ponti deducibili dall'interno di un regime inferiore senza assunzioni aggiuntive sulla struttura causale del dominio.

Il primo livello, quello dell'**associazione**, è il regno della statistica osservazionale. A questo livello, l'agente — umano o artificiale — registra le co-occorrenze nelle distribuzioni dei dati, condiziona sulla presenza di variabili osservate e genera previsioni della forma $P(Y \mid X = x)$: la probabilità che l'evento Y si verifichi dato che si è osservato $X = x$. La competenza associativa è preziosa e sufficiente per un ampio spettro di compiti pratici — riconoscimento di pattern, classificazione, predizione a breve termine in ambienti stazionari — ma è intrinsecamente cieca alla struttura che genera i dati osservati. Un agente che opera esclusivamente a questo livello non può rispondere a domande del tipo "cosa succederebbe se cambiassi X ?" perché questo tipo di domanda presuppone un'operazione che non è definibile all'interno della distribuzione osservazionale congiunta.

Il secondo livello, quello dell'**intervento**, è il dominio delle distribuzioni interventistiche $P(Y \mid \text{do}(X = x))$, dove l'operatore $\text{do}(\cdot)$ — introdotto formalmente da Pearl nella notazione del calcolo degli interventi — denota l'imposizione esogena di un valore alla variabile X , indipendentemente da qualsiasi meccanismo che altrimenti la determinerebbe. La distribuzione interventistica $P(Y \mid \text{do}(X = x))$ è in generale diversa dalla distribuzione condizionale osservazionale $P(Y \mid X = x)$ ogni volta che X è causalmente influenzata da variabili che confondono la relazione tra X e Y — ovvero in presenza di confounders non osservati. L'identità tra le due distribuzioni si verifica soltanto in condizioni speciali: quando X è esogenamente assegnata

(come in un esperimento randomizzato), o quando il grafo causale soddisfa le condizioni che rendono applicabile il backdoor criterion o il frontdoor criterion.

Il terzo livello, quello del **controfattuale**, è il dominio delle proposizioni del tipo $P(Y_{\neg x} = y' \mid X = x, Y = y)$: la probabilità che Y avrebbe assunto il valore y' se X fosse stato x , dato che nel mondo effettivo si è osservato $X = x$ e $Y = y$. Il ragionamento controfattuale richiede di confrontare il mondo effettivo con un mondo possibile alternativo in cui la variabile X è stata manipolata in modo controfattuale — un'operazione che presuppone non soltanto un modello causale strutturale (Structural Causal Model, SCM) del dominio, ma anche la specificazione di funzioni causali deterministe (o stocastiche) per ciascuna variabile del modello.

La distinzione fondamentale che questa gerarchia impone all'analisi dei sistemi di IA è la seguente: qualsiasi sistema addestrato esclusivamente su dati osservazionali — sia esso un LLM addestrato su testo scritto, un sistema JEPA addestrato su flussi video, o un agente reinforcement learning che ha osservato traiettorie altrui senza intervenire nell'ambiente — è in linea di principio confinato al primo livello della gerarchia. Accedere al secondo livello richiede o un'interazione interventistica diretta con l'ambiente, o un modello causale strutturale del dominio che sia stato acquisito per altra via. Accedere al terzo livello richiede entrambe le condizioni più la capacità di ragionare su controfattuali — un'operazione che nella logica standard non è definibile senza un'ontologia modale, e che nella formalizzazione di Pearl richiede la specifica completa dello SCM.

6.2 Il calcolo degli interventi e il problema dell'identificabilità

La formalizzazione precisa di quando una distribuzione interventistica $P(Y \mid \text{do}(X))$ può essere calcolata dalla distribuzione osservazionale $P(V)$ — dove V è l'insieme di tutte le variabili osservate — è il problema dell'**identificabilità causale**, risolto da Pearl attraverso l'algoritmo do-calculus. Il do-calculus consiste in tre regole di trasformazione che permettono di manipolare espressioni contenenti l'operatore $\text{do}(\cdot)$:

Regola 1 (Inserzione/Cancellazione di Osservazione):

$$P(Y \mid \text{do}(X), Z, W) = P(Y \mid \text{do}(X), W)$$

se Z è d-separato da Y dato $\{X, W\}$ nel grafo causale $G_{\neg X}$ in cui sono stati rimossi tutti gli archi entranti in X (il grafo in cui X è "tagliata" dalla sua causa). Questa regola afferma che se, una volta che si interviene su X , la variabile Z diventa indipendente da Y condizionatamente a W , allora l'osservazione di Z non aggiunge informazione su Y .

Regola 2 (Scambio Azione/Osservazione):

$$P(Y \mid \text{do}(X), \text{do}(Z), W) = P(Y \mid \text{do}(X), Z, W)$$

se Z è d-separato da Y dato $\{X, W\}$ nel grafo $G_{\underline{X}Z}$ in cui sono stati rimossi gli archi entranti in X e gli archi uscenti da Z . Questa regola permette di sostituire un intervento su Z con un'osservazione di Z quando le due operazioni sono causalmente equivalenti.

Regola 3 (Inserzione/Cancellazione di Azione):

$$P(Y \mid \text{do}(X), \text{do}(Z), W) = P(Y \mid \text{do}(X), W)$$

se Z è d-separato da Y dato $\{X, W\}$ nel grafo $G_{\underline{X}, Z(W)}$ in cui sono stati rimossi gli archi entranti in X e gli archi entranti in Z che non sono predecessori di W . Questa regola permette di eliminare un intervento su Z quando Z non ha effetto causale su Y una volta che si condiziona su W .

Pearl ha dimostrato che il do-calculus è **completo**: una distribuzione interventistica è identificabile dalla distribuzione osservazionale se e soltanto se può essere trasformata in un'espressione puramente osservazionale attraverso l'applicazione ripetuta di queste tre regole. L'algoritmo ID di Shpitser e Pearl (2006) e le sue generalizzazioni forniscono un procedimento sistematico per verificare l'identificabilità e, quando essa sussiste, calcolare l'espressione osservazionale equivalente.

Le implicazioni per i sistemi di apprendimento automatico sono di portata notevole. Un modello che apprende soltanto dalla distribuzione osservazionale non apprende implicitamente il grafo causale sottostante: apprende la distribuzione congiunta $P(V)$, che in generale è compatibile con molteplici grafi causali distinti. Questo fenomeno — la non-identificabilità del grafo causale dalla sola distribuzione osservazionale — è noto come il problema della **Markov equivalence**: due grafi causali con la stessa struttura di d-separazione producono la stessa distribuzione osservazionale ma prevedono distribuzioni interventistiche diverse. Senza interazione interventistica con l'ambiente, o senza la specifica di assunzioni a priori sulla struttura causale, non è possibile distinguere tra modelli causalmente equivalenti.

6.3 LLM e causalità: la competenza pseudo-causale da statistica distribuzionale

I modelli linguistici di larga scala manifestano, nelle interazioni con utenti esperti, ciò che potremmo chiamare **competenza pseudo-causale**: la capacità di articolare relazioni causali corrette — "il riscaldamento globale è causato dall'accumulo di gas serra nell'atmosfera", "la penicillina uccide i batteri interferendo con la sintesi della parete cellulare" — in modo fluente e apparentemente informato. Questa competenza è il risultato di un processo di acquisizione radicalmente diverso da quello che un agente che ha costruito

internamente un modello causale del dominio metterebbe in atto, e la distinzione tra le due modalità di acquisizione ha conseguenze dirette sulla robustezza e la generalizzabilità della conoscenza causale dei due tipi di sistemi.

Il modello linguistico acquisisce le relazioni causali come **pattern statistici nel corpus testuale**: le parole "causa", "produce", "genera", "determina", e le strutture sintattiche corrispondenti co-occorrono con specifiche coppie di variabili in modo sistematico, e il modello apprende queste co-occorrenze come parte della distribuzione del linguaggio naturale. La rappresentazione interna che il modello costruisce di una relazione causale non è — strutturalmente non può essere — una struttura che codifichi il meccanismo causale sottostante: è una struttura che codifica la distribuzione delle espressioni linguistiche che descrivono quella relazione causale.

Questa differenza ha conseguenze pratiche identificabili. Innanzitutto, la competenza causale degli LLM è **correlata con la frequenza nel corpus**: relazioni causali di cui si parla frequentemente nella letteratura disponibile vengono articolate correttamente con più alta probabilità rispetto a relazioni meno documentate, indipendentemente dalla complessità o dalla certezza della relazione. In secondo luogo, la competenza causale degli LLM è **fragile rispetto al cambiamento di contesto**: quando la stessa relazione causale viene presentata in un contesto linguistico inusuale, o quando si chiede al modello di ragionare su un controfattuale che non è direttamente documentato nel corpus, le prestazioni degradano in modo non prevedibile. In terzo luogo — e questa è la conseguenza più profonda — gli LLM non possono in linea di principio distinguere tra le tre direzioni causali possibili tra due variabili correlate X e Y : $X \rightarrow Y$, $Y \rightarrow X$, e $X \leftarrow U \rightarrow Y$ (con U confonditore latente). La distribuzione testuale descrive le correlazioni, e le correlazioni — come Pearl ha dimostrato con il suo teorema fondamentale — sono simmetriche rispetto alle direzioni causali.

Un esperimento mentale illuminante: si consideri un LLM addestrato su un corpus che contenesse sistematicamente la descrizione di una correlazione spuria — diciamo, la correlazione documentata tra le vendite di gelati e il numero di annegamenti nelle piscine negli Stati Uniti, entrambe variabili confuse dalla temperatura estiva. Un LLM addestrato su tale corpus articolerebbe probabilmente la relazione causale nella direzione della correlazione osservata, inferendo un nesso causale tra gelati e annegamenti che non esiste. Non avrebbe la capacità strutturale di riconoscere che la correlazione è spuria senza informazione aggiuntiva sulla struttura causale del dominio — informazione che non è deducibile dalla distribuzione osservazionale del corpus.

La tecnica del **chain-of-thought prompting** (Wei et al., 2022) e le sue varianti — zero-shot chain-of-thought, least-to-most prompting, tree of

thoughts — rappresentano tentativi pragmatici di migliorare il ragionamento causale degli LLM decomponendo i problemi in passi intermedi. L'efficacia di queste tecniche è reale e documentata su benchmark di ragionamento multi-passo, ma la loro interpretazione è dibattuta: esse migliorano le prestazioni su compiti in-distribution, ma il grado in cui attivano genuino ragionamento causale piuttosto che un'imitazione statistica più sofisticata del ragionamento causale — quello che Ullman (2023) ha chiamato "reasoning that looks like reasoning" — rimane aperto.

Formalmente, la differenza può essere espressa nel modo seguente. Un sistema con genuina comprensione causale del dominio possiede una rappresentazione del modello causale strutturale $M = (V, U, F, P_U)$, dove V è l'insieme delle variabili endogene, U è l'insieme delle variabili esogene (rumore), F è l'insieme delle funzioni causali strutturali $f_i : V_i \times U_i \rightarrow V_i$ che determinano il valore di ciascuna variabile in funzione dei suoi genitori causali e del rumore, e P_U è la distribuzione delle variabili esogene. Questo sistema può rispondere a domande interventistiche calcolando $P(Y \mid \text{do}(X = x))$ come segue:

sia M_x il modello modificato per intervento in cui la funzione f_X è sostituita dalla funzione costante $f_X(\cdot) = x$. Allora:

$$P(Y \mid \text{do}(X = x)) = P_{\{M_x\}}(Y)$$

ovvero la distribuzione di Y nel modello modificato. Un LLM non possiede questa struttura: possiede una distribuzione $P(Y_t \mid X_{\leq t})$ sul prossimo token condizionato ai token precedenti, che è una distribuzione osservazionale sul corpus testuale, non una distribuzione su interventi nel mondo.

6.4 JEPA e la causalità: la via dell'intervento incarnato

Le architetture JEPA nel framework A-B-M di Dupoux, LeCun e Malik, e specificamente la componente del Sistema B — l'apprendimento per azione — sono progettate per accedere, almeno parzialmente, al secondo livello della gerarchia causale di Pearl attraverso l'interazione interventistica con l'ambiente. Questo accesso, tuttavia, non è automatico né garantito dalla sola presenza di un meccanismo d'azione: richiede che l'azione soddisfi precise condizioni strutturali che i modelli A-B-M attuali non sempre realizzano, e la cui formalizzazione è uno dei contributi concettuali più importanti che la teoria causale può offrire al campo dell'intelligenza artificiale incarnata.

Il punto di partenza formale è la distinzione, introdotta da Peters, Mooij, Janzing e Schölkopf in *Elements of Causal Inference* (2017), tra la distribuzione osservazionale $P(S_{t+1} \mid S_t, A_t)$ — dove S_t è lo stato al tempo t e A_t è l'azione eseguita — e la distribuzione interventistica $P(S_{t+1} \mid \text{do}(A_t = a))$. La differenza tra le due distribuzioni dipende dalla struttura causale del

processo che determina le azioni. Se le azioni sono generate da una policy $\pi(A_t | S_t)$ che è una funzione dello stato, allora le azioni e gli stati sono causalmente dipendenti, e la distribuzione condizionale $P(S_{t+1} | S_t, A_t)$ che si osserva durante l'interazione autonoma dell'agente confonde la variazione causale dell'azione con la selezione delle azioni in base allo stato. Formalmente:

$$P(S_{t+1} | S_t, A_t) = \sum_{S'_t} P(S_{t+1} | S'_t, A_t) P(S'_t | S_t, A_t)$$

dove l'ultimo termine, $P(S'_t | S_t, A_t)$, non è necessariamente identico a $P(S'_t | S_t)$ se l'azione A_t è correlata con caratteristiche dello stato non completamente osservate. In presenza di confounders latenti tra azione e stato, la distribuzione condizionale dell'azione sullo stato è distorta, e il modello del mondo appreso dal Sistema A su dati generati da tale policy eredita questa distorsione.

La soluzione teoricamente corretta, analoga al design sperimentale nelle scienze empiriche, è l'**esplorazione randomizzata**: se le azioni vengono selezionate da una distribuzione uniforme (o da qualsiasi distribuzione che sia indipendente dallo stato), allora l'agente esegue di fatto un esperimento randomizzato nell'ambiente, e la distribuzione delle transizioni osservate converge alla distribuzione interventistica. I meccanismi di curiosità e di esplorazione intrinseca del Sistema B — discussi nel paper di Dupoux et al. nella sezione 2.3 — possono essere interpretati, in questa luce, non soltanto come strategie per ottimizzare la copertura dello spazio degli stati, ma come meccanismi per garantire la variazione quasi-sperimentale delle azioni necessaria per l'inferenza causale robusta.

Questa interpretazione ha un precedente diretto nella letteratura di psicologia dello sviluppo che Dupoux et al. citano: Gopnik e colleghi hanno documentato che i bambini in età prescolare eseguono esperimenti sistematici sull'ambiente fisico e sociale — variando un fattore alla volta, osservando l'effetto, e aggiornando il loro modello causale — in modo strutturalmente analogo al design di Millian dei metodi di concordanza e differenza. Il bambino che inserisce il cubo nel foro quadrato e non nel foro circolare sta eseguendo un intervento sull'ambiente ($\text{do}(\text{forma} = \text{quadrato})$) e osservando l'effetto (il cubo entra), separando causalmente la forma dell'oggetto dalla forma del foro in modo che la semplice osservazione — vedere qualcuno inserire il cubo — non consentirebbe con la stessa chiarezza.

La componente JEPA del Sistema A, nel framework A-B-M, apprende un modello predittivo delle transizioni nello spazio latente:

$$\hat{z}_{t+1} = f_{\theta}(z_t, a_t)$$

dove $z_t = g_\varphi(s_t)$ è la rappresentazione latente dello stato s_t e a_t è l'azione. Questo modello è addestrato minimizzando la perdita di previsione:

$$L_{\text{pred}} = E_{(s_t, a_t, s_{t+1}) \sim D} [d(f_\theta(g_\varphi(s_t), a_t), g_\varphi(s_{t+1}))]$$

dove $d(\cdot, \cdot)$ è una misura di distanza nello spazio latente e D è la distribuzione dei dati di addestramento. La questione causale fondamentale è: quanto la coppia (f_θ, g_φ) ottimale per questa perdita approssima la struttura causale del mondo — ovvero, in che misura l'encoder g_φ apprende rappresentazioni che corrispondono a variabili causalmente rilevanti, e in che misura il predittore f_θ apprende meccanismi causali piuttosto che correlazioni statistiche nelle traiettorie di addestramento?

La risposta dipende criticamente dalla natura dei dati di addestramento D . Se D è generato da un agente che esplora in modo sistematicamente randomizzato, ovvero se $A_t \perp S_t \mid (S_{t-1}, \dots, S_0)$ — se le azioni sono condizionatamente indipendenti dallo stato corrente dati gli stati passati — allora la distribuzione delle transizioni converge alla distribuzione interventistica, e il modello appreso (f_θ, g_φ) approssima la struttura causale. Se invece D è generato da una policy determinista o quasi-determinata dallo stato — come accade nella maggior parte dei dati di interazione robotica registrati da teleoperation — allora il modello appreso è distorto nella direzione della correlazione osservazionale, e può fallire in modo sistematico quando viene dispiegato in un contesto in cui la struttura causale è la stessa ma la distribuzione degli stati d'ingresso è diversa.

6.5 Rappresentazione causale latente: il programma di Schölkopf e i suoi rapporti con JEPA

Un filone di ricerca che ha guadagnato rilevanza nell'ultimo decennio e che è direttamente pertinente alla valutazione delle architetture JEPA come fondamenti per il ragionamento causale è il programma di **Causally-Inspired Machine Learning** sviluppato da Bernhard Schölkopf e collaboratori, sintetizzato in Schölkopf et al. (2021) e elaborato nel concetto di Causal Representation Learning (Schölkopf et al., 2022; Von Kügelgen et al., 2023). L'obiettivo centrale di questo programma è identificare le condizioni sotto le quali algoritmi di apprendimento della rappresentazione — inclusi gli algoritmi di previsione nello spazio latente di tipo JEPA — apprendono rappresentazioni che corrispondono a variabili causalmente rilevanti del processo generativo dei dati, piuttosto che a correlazioni statistiche arbitrarie.

La tesi fondamentale del programma è che la **robustezza alla distributional shift** — la capacità di generalizzare a distribuzioni di test diverse da quelle di addestramento — è strettamente correlata con l'acquisizione di rappresentazioni causali piuttosto che puramente predittive. Formalmente, il

risultato chiave è il seguente: se le variabili latenti che un sistema di rappresentazione apprende corrispondono ai meccanismi causali indipendenti che generano i dati osservabili — ovvero se il sistema ha appreso qualcosa di analogo ai "fattori causali" del processo generativo — allora il sistema è in grado di generalizzare a distribuzioni generate da interventi su singoli meccanismi, perché gli altri meccanismi rimangono invarianti sotto tali interventi.

Questo risultato formale — la **Generalization via Causal Mechanism** di Schölkopf — può essere enunciato in modo preciso. Si consideri un processo generativo dei dati $X = (X_1, \dots, X_n)$ con struttura causale specificata da uno SCM $M = (V, U, F, P_U)$. Si assuma che i dati di addestramento siano generati da M , e che le distribuzioni di test siano generate da versioni perturbate M' in cui un sottoinsieme $S \subseteq V$ dei meccanismi causali è stato modificato (interventi o soft interventions). Un sistema di rappresentazione che ha appreso variabili latenti corrispondenti ai meccanismi causali individuali — e non a combinazioni statisticamente convenienti ma causalmente eterogenee di tali meccanismi — è in grado di generalizzare alla distribuzione di test modificando soltanto le componenti della propria rappresentazione corrispondenti ai meccanismi in S , mentre le componenti corrispondenti ai meccanismi non modificati rimangono stabili.

La domanda pertinente per la valutazione di JEPA è se l'obiettivo di previsione nello spazio latente — la minimizzazione di $d(f_\theta(g_\varphi(s_t), a_t), g_\varphi(s_{t+1}))$ — produce una pressione di selezione sulla rappresentazione $g_\varphi(\cdot)$ che favorisce l'acquisizione di variabili causalmente rilevanti rispetto a variabili meramente correlative. La risposta teorica è **parzialmente affermativa** per ragioni che vale la pena esaminare in dettaglio.

L'architettura JEPA — e questo è uno dei suoi vantaggi strutturali rispetto ai modelli generativi ricostruttivi — impone al predittore f_θ di operare nello spazio latente piuttosto che nello spazio delle osservazioni. Ciò significa che il predittore deve trovare una rappresentazione $z_t = g_\varphi(s_t)$ in cui la variazione prodotta dall'azione a_t sia prevedibile — ovvero in cui le variabili latenti siano sensibili alle conseguenze delle azioni, non ai dettagli percettivi irrilevanti per il cambiamento causato dall'azione. Questa pressione di selezione sulla rappresentazione è analoga alla pressione che nella psicologia dello sviluppo porta i bambini ad astrarre dagli attributi percettivi superficiali degli oggetti per rappresentarli attraverso le loro proprietà funzionali — le affordance nel senso di Gibson.

Tuttavia, questa pressione non è sufficiente, da sola, a garantire che le variabili latenti corrispondano a variabili causalmente interpretabili nel senso di Schölkopf. Rimangono due problemi fondamentali. Il primo è il problema dell'identificabilità **della rappresentazione causale**: anche assumendo che i dati di addestramento siano generati da un processo con struttura causale

nota, esistono in generale molteplici sistemi di variabili latenti che minimizzano ugualmente bene la perdita di previsione nello spazio latente, e la maggior parte di questi sistemi non corrisponde alle variabili causalmente rilevanti del processo generativo. I risultati di Von Kügelgen et al. (2023) mostrano che l'identificabilità della rappresentazione causale richiede assunzioni di non-linearità e di struttura del rumore che non sono in generale soddisfatte, e che le architetture JEPA attuali non impongono esplicitamente.

Il secondo problema è la **spurious feature correlation**: durante l'addestramento, variabili latenti che co-variano con i cambiamenti causalmente prodotti dalle azioni possono essere apprese anche se non sono causalmente rilevanti — se sono soltanto correlate con le variabili causalmente rilevanti nel dataset di addestramento. Questo fenomeno è analogo al shortcut learning documentato da Geirhos et al. (2020) per le reti convoluzionali: i modelli apprendono caratteristiche correlate con l'etichetta target nel dataset di addestramento, ma che non sono causalmente legate alla classe dell'oggetto. Il risultato è un modello che performa bene in-distribution ma generalizza male fuori dalla distribuzione di addestramento — esattamente la forma di fragilità che il programma di Causal Representation Learning mira ad eliminare.

6.6 Il controfattuale come frontiera: ragionamento immaginativo e modelli del mondo

Il terzo livello della gerarchia causale di Pearl — il ragionamento controfattuale — è quello in cui la distanza tra i sistemi di intelligenza artificiale attuali e le capacità cognitive umane è più ampia e più concettualmente profonda. Il ragionamento controfattuale richiede la capacità di costruire e valutare scenari alternativi al mondo effettivo: "se avessi preso la strada diversa, sarei arrivato prima?", "se il paziente non avesse fumato, avrebbe contratto il cancro ai polmoni?", "se il robot avesse afferrato l'oggetto diversamente, avrebbe evitato di farlo cadere?". Queste domande non riguardano distribuzioni di eventi possibili — sono domande sul singolo caso controfattuale, che richiede l'individuazione del meccanismo causale che connette la manipolazione ipotetica all'esito alternativo.

La formalizzazione di Pearl dei controfattuali procede in tre passi che costituiscono l'algoritmo **Abduction-Action-Prediction** (AAP):

Passo 1 — Abduzione: dato l'evidenza osservato ($S_t = s, A_t = a, S_{t+1} = s'$), inferire la distribuzione delle variabili esogene U condizionatamente all'osservazione:

$$P(U \mid S_t = s, A_t = a, S_{t+1} = s')$$

Questo passo richiede di "invertire" il modello causale strutturale per dedurre i valori del rumore esogeno che sono consistenti con l'osservazione.

Passo 2 — Azione: modificare il modello causale M in un modello controfattuale $M_{\{a'\}}$ sostituendo la funzione causale f_A con la costante a' — il valore controfattuale ipotetico dell'azione.

Passo 3 — Previsione: calcolare la distribuzione dell'esito Y nel modello controfattuale $M_{\{a'\}}$ con i valori delle variabili esogene inferiti nel passo 1:

$$P(Y_{\{a'\}} = y' \mid S_t = s, A_t = a, S_{\{t+1\}} = s')$$

Il risultato è la distribuzione dell'esito nel mondo controfattuale — nel mondo in cui l'azione fosse stata a' invece di a , mantenendo tutto il resto uguale al mondo effettivo.

Il papel di Dupoux et al. accenna a questa capacità nella sezione sull'apprendimento per immaginazione (Appendice B.2) e nella discussione del planning multi-passo: un sistema con un modello del mondo può "immaginare" traiettorie controfattuali nello spazio latente, eseguendo il predittore f_θ su sequenze di azioni ipotetiche partendo dalla rappresentazione latente dello stato corrente. Formalmente:

$$\hat{z}_{\{t+k\}} = f_\theta^k(z_t, a'_{\{t:t+k\}})$$

dove f_θ^k denota l'applicazione iterata k volte del predittore con la sequenza di azioni ipotetiche $a'_{\{t:t+k\}}$. Questa operazione è strutturalmente analoga al secondo e terzo passo dell'algoritmo AAP di Pearl, con la rappresentazione latente z_t che svolge il ruolo dell'abduzione delle variabili esogene.

Tuttavia, la struttura di questa operazione di "immaginazione" differisce dal controfattuale Pearleano in un aspetto cruciale: il passo di abduzione — l'inferenza delle variabili esogene consistenti con l'osservazione effettiva — non è eseguito esplicitamente. Il sistema JEPA, quando esegue il roll-out latente da z_t con azioni ipotetiche, non condiziona su ciò che è effettivamente accaduto nel mondo: si limita a propagare in avanti la rappresentazione latente dello stato corrente attraverso il predittore condizionato sulle azioni ipotetiche. Questo equivale al passo 3 dell'algoritmo AAP senza il passo 1: una previsione del futuro sotto azioni diverse, non un ragionamento controfattuale in senso stretto.

La differenza pratica tra le due operazioni emerge quando le variabili esogene U — i fattori che nel mondo effettivo non erano controllati dall'agente — sono causalmente rilevanti per l'esito. Un sistema che esegue il passo 1 (abduzione) inferisce i valori di U consistenti con l'osservazione e condiziona il controfattuale su questi valori, ottenendo una previsione dell'esito nel mondo alternativo che è coerente con ciò che si sa del mondo effettivo. Un sistema che salta il passo 1 — come il roll-out latente JEPA — ottiene una previsione che è coerente con la distribuzione media delle variabili esogene nel dataset di addestramento, non con i valori specifici che U aveva nel caso

effettivo. La differenza è la differenza tra "cosa succederebbe, in media, se eseguiassi l'azione a' ?" e "cosa sarebbe successo in questo caso specifico se avessi eseguito l'azione a' invece di a ?". La prima è una previsione interventistica (livello 2 di Pearl), la seconda è un controfattuale genuino (livello 3).

6.7 Invarianza causale e la robustezza al distributional shift: un ponte tra teoria e pratica

La connessione tra causalità e robustezza alla distributional shift — identificata come uno dei problemi centrali nella pratica dell'IA da Geirhos et al. (2020) e da Koh et al. (2021) — è il punto in cui le considerazioni teoriche della causalità si traducono più direttamente in metriche di valutazione pratica. L'**Invariant Causal Prediction** (ICP) di Peters, Bühlmann e Meinshausen (2016) e i metodi correlati — **Invariant Risk Minimization** (IRM) di Arjovsky et al. (2019), **Risk Extrapolation** (REx) di Krueger et al. (2021) — formalizzano la connessione tra rappresentazione causale e invarianza della prestazione attraverso distribuzioni di test.

Il principio fondamentale è il seguente. Si consideri un sistema che deve prevedere Y a partire da X in molteplici ambienti $\{E_1, \dots, E_n\}$, dove gli ambienti differiscono per la distribuzione di alcune variabili di intervento. I meccanismi causali che connettono le cause di Y a Y sono **invarianti** attraverso gli ambienti — la struttura causale è la stessa, ciò che cambia è la distribuzione delle variabili esogene che perturbano l'ambiente. Un sistema che ha appreso le variabili causalmente rilevanti per Y avrà prestazioni stabili attraverso gli ambienti; un sistema che ha appreso correlazioni spurie — che sono stabili in alcuni ambienti ma non in tutti — avrà prestazioni degradate negli ambienti di test che non erano presenti nell'addestramento.

L'obiettivo di IRM è trovare una rappresentazione $\varphi(X)$ e un classificatore $w \circ \varphi(X)$ tale che:

$$\operatorname{argmin}_{\{w, \varphi\}} \sum_e R^e(w \circ \varphi) \text{ soggetto a: } w \in \operatorname{argmin}_{\{\tilde{w}\}} R^e(\tilde{w} \circ \varphi) \text{ per ogni } e \in E_{\text{train}}$$

dove R^e è il rischio nell'ambiente e . La condizione di vincolo richiede che il classificatore lineare w sia ottimale simultaneamente per tutti gli ambienti di addestramento — il che è possibile soltanto se $\varphi(X)$ ha estratto le componenti di X che sono causalmente stabili attraverso gli ambienti.

Applicato alle architetture JEPA, questo framework suggerisce un modo di valutare empiricamente se la rappresentazione latente $g_\varphi(\cdot)$ ha acquisito struttura causale: la rappresentazione è causale nella misura in cui un predittore lineare addestrato su di essa su dati di un ambiente E_1 mantiene le proprie prestazioni su dati di un ambiente E_2 che differisce per la distribuzione di variabili correlate in modo spurio con il target. I benchmark

di fisica intuitiva — come IntPhys (Riochet et al., 2018) e CATER (Girdhar e Ramanan, 2020) — sono progettati precisamente per valutare questa dimensione: testare se i sistemi mantengono la comprensione delle proprietà fisiche degli oggetti (permanenza, continuità, solidità) quando la distribuzione dei dettagli percettivi irrilevanti cambia.

I risultati di Garrido et al. (2025) — citati nel paper di Dupoux et al. — mostrano che V-JEPA pre-addestrato su video naturali sviluppa comprensione intuitiva della fisica nel senso di IntPhys significativamente superiore ai modelli ricostruttivi e agli LLM multimodali confrontabili, il che è coerente con la tesi che la previsione nello spazio latente produca una pressione di selezione verso rappresentazioni causalmente più robuste. Tuttavia, come discusso nella sezione precedente, questa coerenza non è ancora conclusiva: V-JEPA potrebbe avere acquisito correlazioni statistiche fisicamente strutturate piuttosto che meccanismi causali nel senso formale di Schölkopf, e la distinzione tra le due ipotesi richiederebbe esperimenti di intervento controllato che non sono stati ancora sistematicamente condotti.

6.8 Verso una teoria causale dell'apprendimento per osservazione e per azione

Il contributo concettuale più rilevante che la cornice Pearleana offre al programma di ricerca di Dupoux, LeCun e Malik è la possibilità di formalizzare con precisione ciò che i due sistemi fondamentali del framework A-B-M — il Sistema A (osservazione) e il Sistema B (azione) — possono e non possono apprendere riguardo alla struttura causale del mondo, e in che modo la loro interazione può ampliare le capacità causali del sistema rispetto a ciascuno dei due componenti in isolamento.

Il Sistema A, operando su dati osservazionali, è confinato al primo livello della gerarchia causale: può apprendere le distribuzioni marginali e condizionali $P(V)$, comprese le dipendenze condizionali che corrispondono all'assenza di archi di d-separazione nel grafo causale, ma non può distinguere tra direzioni causali alternative né inferire le distribuzioni interventistiche. Nella terminologia del do-calculus, il Sistema A può apprendere le espressioni $P(Y \mid X)$ ma non le espressioni $P(Y \mid \text{do}(X))$.

Il Sistema B, attraverso l'interazione interventistica con l'ambiente, introduce variazione quasi-sperimentale nelle azioni: le azioni esplorative che soddisfano la condizione di indipendenza dall'osservazione producono dati della distribuzione interventistica $P(S_{t+1} \mid \text{do}(A_t = a))$, permettendo al Sistema A di aggiornare il proprio modello verso la distribuzione interventistica corretta. In questo senso, il Sistema B svolge il ruolo metodologico dello sperimentatore che arricchisce il dataset osservazionale del Sistema A con dati che consentono l'identificazione causale.

La sinergia tra i due sistemi, orchestrata dal Sistema M, può in linea di principio condurre a un sistema che si avvicina al secondo livello della gerarchia causale — un sistema con un modello del mondo approssimativamente interventistico. Questo è il nucleo teorico più robusto del framework A-B-M, e la sua formalizzazione nel linguaggio della teoria causale di Pearl è un contributo che il paper di Dupoux et al. non offre esplicitamente, ma che è implicito nella struttura della proposta.

L'accesso al terzo livello — il ragionamento controfattuale — rimane strutturalmente più problematico. Come argomentato nella sezione precedente, il roll-out latente JEPA realizza la previsione interventistica (livello 2) ma non il controfattuale Pearleano (livello 3) nella sua forma completa, perché manca il passo di abduzione che condiziona il controfattuale sui valori specifici delle variabili esogene nell'istanza effettiva. Architetture che eseguano esplicitamente questo passo — inferendo le variabili latenti esogene consistenti con l'osservazione e poi propagando in avanti il modello con azioni alternative — avrebbero una capacità di ragionamento controfattuale strutturalmente più ricca, e la loro implementazione rappresenta una delle direzioni di ricerca più promettenti nell'integrazione tra causalità e apprendimento per rappresentazione.

La distanza tra la competenza causale dei sistemi attuali e quella degli organismi biologici rimane, al termine di questa analisi, considerevole. I bambini di cinque anni che Gopnik et al. hanno documentato come "piccoli scienziati causali" eseguono esperimenti sistematici, inferiscono strutture causali, ragionano controfattualmente e generalizzano queste strutture a nuovi contesti con una facilità e una robustezza che nessun sistema attuale — né LLM né World Model JEPA — ha ancora replicato. Il programma A-B-M è il tentativo più ambizioso e teoricamente fondato per colmare questa distanza, e la sua formalizzazione nel linguaggio della teoria causale è uno degli sviluppi che ne determinerà, in ultima analisi, il successo o l'insuccesso come programma scientifico unificato.

Sezione VII — Formalizzazione del Sistema M: Meta-Controllo, Orchestrazione Cognitiva e il Problema dell'Autoorganizzazione

7.1 Il Sistema M come piano di controllo distribuito: l'analogia con le reti definite dal software

Il paper di Dupoux, LeCun e Malik introduce il Sistema M — il meta-controllore — con un'analogia che vale la pena sviluppare fino alle sue implicazioni computazionali più profonde: il confronto con il Control Plane delle Software-Defined Networks (SDN). Nelle architetture di rete definite dal software, il Control Plane — che gestisce le decisioni di routing, la

configurazione delle tabelle di forwarding e la topologia della rete — è separato dal Data Plane, che gestisce il flusso effettivo dei pacchetti. Il Control Plane opera su una rappresentazione astratta della rete — la sua topologia, i carichi di traffico, le politiche di qualità del servizio — e genera configurazioni che il Data Plane implementa. Questa separazione consente al Control Plane di reagire a cambiamenti globali della rete con agilità impossibile in architetture non-SDN, dove ogni nodo deve scoprire localmente la topologia e negoziare il routing con i vicini.

L'analogia con il Sistema M è strutturalmente precisa: il Sistema M non elabora i flussi di dati ad alta frequenza del mondo sensoriale (le immagini, i suoni, le sequenze motorie) — questo è il compito del Sistema A e del Sistema B. Il Sistema M opera su una rappresentazione di basso ingresso-dati e alta interpretazione semantica del funzionamento dei due sistemi subordinati: i segnali di meta-stato (prediction error, uncertainty, epistemic signals, somatic signals) che il paper classifica in tre categorie — epistemici, specifici della specie e somatici. Il Sistema M genera meta-azioni che riconfigurano dinamicamente le connessioni tra i componenti del sistema: apre e chiude i canali di flusso dei dati, commuta i componenti tra modalità di apprendimento e modalità di inferenza, accede alla memoria episodica per replay, modula le funzioni di perdita dei sistemi subordinati.

Per formalizzare questa struttura in modo matematicamente rigoroso è necessario un linguaggio formale che catturi tre dimensioni del Sistema M: la sua struttura di stato (lo spazio dei meta-stati), la sua struttura di azione (lo spazio delle meta-azioni), e la sua politica (la funzione che mappa i meta-stati sulle meta-azioni). Il paper di Dupoux et al. abbozza questi tre elementi in modo schematico — la Figura 4 del paper e la discussione della sezione 3 — ma senza la precisione formale che sarebbe necessaria per un'implementazione computazionale. La presente sezione si propone di colmare questa lacuna.

7.2 Spazio dei meta-stati: una formalizzazione tripartita

Il paper classifica i meta-stati in tre categorie: epistemici, specifici della specie e somatici. Proponiamo di formalizzare questa classificazione come una struttura di prodotto cartesiano sullo spazio totale dei meta-stati S^M :

$$S^M = S^M_{\epsilon} \times S^M_{\sigma} \times S^M_{\omega}$$

dove S^M_{ϵ} è lo spazio dei meta-stati epistemici, S^M_{σ} è lo spazio dei meta-stati specifici della specie, e S^M_{ω} è lo spazio dei meta-stati somatici. Ciascuno di questi sottospazi è a sua volta strutturato.

Lo spazio dei meta-stati epistemici S^M_{ϵ} cattura le proprietà del funzionamento interno dei sistemi A e B che sono rilevanti per le decisioni di

meta-controllo. Proponiamo di rappresentarlo come un vettore di dimensione d_ϵ i cui componenti includono:

$$s^\epsilon = (\delta_A, \sigma^2_A, \eta_A, \delta_B, \sigma^2_B, \eta_B, \Delta_{AB}, \varrho)$$

dove $\delta_A \in \mathbb{R}^+$ è l'errore di previsione medio del Sistema A — la perdita di previsione normalizzata per la scala dei dati, che il paper identifica come equivalente del prediction error biologico di Schultz et al. (1997); $\sigma^2_A \in \mathbb{R}^+$ è la varianza dell'errore di previsione del Sistema A, come misura dell'incertezza epistemica del modello del mondo; $\eta_A \in \mathbb{R}$ è il tasso di apprendimento del Sistema A, stimabile dalla variazione della perdita tra epoche successive; $\delta_B \in \mathbb{R}$ è la media del reward cumulato del Sistema B — opposta al segno della perdita di controllo — come misura della qualità della policy corrente; $\sigma^2_B \in \mathbb{R}^+$ è la varianza del reward del Sistema B, come misura dell'instabilità dell'ambiente o della policy; $\eta_B \in \mathbb{R}$ è il tasso di aggiornamento della policy del Sistema B; $\Delta_{AB} \in \mathbb{R}$ è la divergenza tra le rappresentazioni latenti del Sistema A e i target del Sistema B — una misura del disallineamento tra il modello del mondo e le esigenze di controllo; $\varrho \in [0,1]$ è il coefficiente di correlazione tra gli errori di previsione del Sistema A e i segnali di reward del Sistema B, come misura del grado di integrazione funzionale tra i due sistemi.

Lo spazio dei meta-stati specifici della specie S^M_σ cattura le proprietà dell'ambiente esterno che, in organismi biologici, sono riconosciute da rivelatori pre-cablati evolutivamente fissi. Per un agente artificiale, questi corrispondono a rivelatori di pattern ad alta priorità che il Sistema M deve monitorare indipendentemente dallo stato dei sistemi A e B. La Tabella C.1 del paper identifica esempi che includono volti e vocalizzazioni, direzione dello sguardo, segnali pedagogici, movimenti biomeccanici auto-propulsi, minacce potenziali (stimoli che si avvicinano rapidamente), e conspecifici dominanti o prestigiosi. Per un agente artificiale in ambienti naturali, la formalizzazione potrebbe includere:

$$s^\sigma = (b_{\text{face}}, b_{\text{motion}}, b_{\text{threat}}, b_{\text{teacher}}, b_{\text{peer}})$$

dove ciascun $b \in \{0, 1\}$ è un segnale binario di rilevamento — o, più realisticamente, $b \in [0,1]$ è un valore di confidenza del rivelatore corrispondente. Questi rivelatori operano con latenza molto bassa rispetto ai processi di apprendimento dei sistemi A e B, e la loro attivazione modifica le priorità delle meta-azioni del Sistema M in modo quasi-riflesso.

Lo spazio dei meta-stati somatici S^M_ω cattura le proprietà dell'agente nella sua interazione con l'ambiente fisico — per un agente incarnato in un corpo fisico. La Tabella C.1 identifica sonno, riposo, dolore, fame e stress come segnali somatici fondamentali. Per un agente robotico, i corrispondenti potrebbero includere:

$$s^\omega = (E_bat, T_motor, F_contact, t_active, \sigma_pose)$$

dove $E_bat \in [0,1]$ è il livello di carica della batteria (analogo all'energia metabolica), $T_motor \in \mathbb{R}^n$ è il vettore delle temperature dei motori (analogo ai segnali di affaticamento muscolare), $F_contact \in \mathbb{R}^6$ è il vettore della forza di contatto nelle sei dimensioni di forza e coppia (analogo al segnale di dolore nella modalità tattile), $t_active \in \mathbb{R}^+$ è il tempo di attivazione continua dall'ultimo episodio di riposo (analogo alla pressione del sonno), e $\sigma_pose \in \mathbb{R}^+$ è l'incertezza sulla posa corrente del robot nell'ambiente (analogo all'orientamento spaziale).

Il meta-stato completo al tempo t è quindi:

$$s^M_t = (s^\varepsilon_t, s^\sigma_t, s^\omega_t) \in S^M$$

La dinamica del meta-stato è determinata dall'evoluzione dei sistemi A e B nell'interazione con l'ambiente:

$$s^M_{t+1} = h(s^M_t, a^M_t, x_{t+1})$$

dove $h(\cdot)$ è la funzione di transizione dei meta-stati, a^M_t è la meta-azione eseguita al tempo t , e x_{t+1} è l'osservazione dell'ambiente al tempo $t+1$. In linea di principio, il Sistema M monitora questa transizione e aggiorna il meta-stato di conseguenza.

7.3 Spazio delle meta-azioni: tipologia e struttura formale

Il paper descrive le meta-azioni come operazioni di routing dinamico sulle connessioni tra i componenti del sistema — l'apertura e la chiusura di canali di flusso dei dati, la commutazione tra modalità operative, la modulazione delle funzioni di perdita. Proponiamo una formalizzazione dello spazio delle meta-azioni A^M come l'insieme delle operazioni che il Sistema M può eseguire sull'architettura:

$$A^M = A^M_route \times A^M_mode \times A^M_loss \times A^M_mem$$

Azioni di routing A^M_route : l'insieme delle possibili connessioni tra i componenti del sistema che il Sistema M può aprire (attivare) o chiudere (disattivare). Formalmente, si consideri un grafo diretto $G = (C, E)$ in cui i nodi $C = \{A, B, M, Mem, Sensor, Motor\}$ rappresentano i componenti del sistema (Sistema A , Sistema B , Sistema M , Memoria Episodica, Sensori, Motori) e gli archi $E \subseteq C \times C \times T$ rappresentano i possibili canali di flusso dei dati con il loro tipo T (dati grezzi, rappresentazioni latenti, policy, reward, meta-stati). Una meta-azione di routing è un sottoinsieme di E che specifica quali canali sono attivi al tempo t :

$$a^{route}_t \in 2^E$$

dove 2^E è l'insieme delle parti di E . Il numero di possibili routing configurations è $2^{|E|}$, che per sistemi non banali è esponenzialmente grande. Il Sistema M deve avere una policy efficiente che restituisca l'azione di routing appropriata per ogni meta-stato senza enumerare esplicitamente tutte le configurazioni possibili.

Azioni di modalità A^M_{mode} : la commutazione di ciascun componente tra modalità di operazione alternative. Per il Sistema A, le modalità rilevanti includono: apprendimento (i pesi vengono aggiornati attraverso la retropropagazione), inferenza (i pesi sono congelati e vengono calcolate le rappresentazioni), campionamento (la memoria episodica viene acceduta per generare batch di replay). Per il Sistema B, le modalità includono: esplorazione (le azioni vengono selezionate per massimizzare l'acquisizione di informazioni, ad esempio attraverso curiosità o randomizzazione), sfruttamento (le azioni vengono selezionate per massimizzare il reward immediato secondo la policy corrente), pianificazione (il Sistema B usa il modello del mondo del Sistema A per simulare traiettorie e scegliere l'azione con il valore simulato più alto). Formalmente:

$$a^{mode_t} = (mode_A, mode_B) \in \{learn, infer, sample\} \times \{explore, exploit, plan\}$$

Azioni di modulazione della perdita A^M_{loss} : la modifica dei parametri delle funzioni di perdita dei sistemi A e B. Per il Sistema A, questo include la modulazione del peso relativo tra diversi obiettivi di previsione (perdita di previsione dello stato futuro vs. perdita di previsione dell'azione vs. perdita di ricostruzione di memoria) e la modulazione del tasso di apprendimento. Per il Sistema B, questo include la modulazione del peso tra reward estrinseco e reward intrinseco (curiosità) e la modifica del discount factor γ che bilancia ricompense a breve e lungo termine:

$$a^{loss_t} = (\lambda_A, \alpha_A, \lambda_B, r^{int_weight}, \gamma_B)$$

dove $\lambda_A \in \mathbb{R}^{\{n_A\}}$ è il vettore dei pesi degli obiettivi del Sistema A, $\alpha_A \in \mathbb{R}^+$ è il tasso di apprendimento del Sistema A, $\lambda_B \in \mathbb{R}^{\{n_B\}}$ è il vettore dei pesi degli obiettivi del Sistema B, $r^{int_weight} \in [0,1]$ è il peso del reward intrinseco rispetto all'estrinseco, e $\gamma_B \in [0,1]$ è il discount factor del Sistema B.

Azioni di memoria A^M_{mem} : le operazioni sull'archivio della memoria episodica, che include la scrittura di nuove esperienze, il recupero di esperienze passate per replay, e la consolidazione di esperienze in memorie a lungo termine. Formalmente:

$$a^{mem_t} = (write_op, read_op, consolidate_op)$$

dove $write_op$ specifica se e quali esperienze del time step corrente vengono scritte nella memoria episodica, $read_op$ specifica il metodo di

campionamento per il replay (uniforme, prioritizzato per errore temporale come nel PER di Schaul et al. (2016), o guidato da curiosità), e `consolidate_op` specifica se viene eseguita una fase di consolidamento offline (equivalente al replay del sonno nella neurobiologia).

7.4 La politica del meta-controllore: struttura formale e vincoli biologici

La politica del Sistema M è una funzione $\pi^M : S^M \rightarrow \Delta(A^M)$ che mappa ogni meta-stato su una distribuzione di probabilità sullo spazio delle meta-azioni. Il paper di Dupoux et al. fa una scelta teorica importante — e controversa — affermando che la politica centrale del Sistema M è "hardwired", ovvero evolutivamente fissa piuttosto che appresa durante la vita dell'agente. Questa scelta merita un'analisi approfondita, perché ha implicazioni dirette sia sulla formalizzazione del Sistema M sia sulla strategia evolutiva per il suo bootstrap.

L'argomentazione a favore di una politica fissa è la seguente: se il Sistema M deve orchestrare l'apprendimento dei sistemi A e B, non può essere esso stesso in apprendimento durante il ciclo di vita dell'agente senza incorrere in una regressione circolare — il problema del "bootstrapping" che il paper identifica come uno dei roadblock principali. Un Sistema M che apprende la propria politica ha bisogno di un meta-meta-controllore che orchestri questo apprendimento, e così via all'infinito. La soluzione proposta è di separare nettamente le scale temporali: il Sistema M è fisso durante il ciclo di vita (scala dello sviluppo), ma la sua struttura è ottimizzata alla scala evolutiva (scala evolutiva) attraverso l'ottimizzazione bi-livello descritta nella sezione 4 del paper.

Questa soluzione è biologicamente plausibile: le strutture prefrontali che svolgono funzioni di meta-controllo negli organismi — il circuito prefrontal-striatal, i circuiti di supervisione dell'attenzione di Shallice e Norman (1986) — mostrano plasticità limitata nella vita adulta rispetto alle strutture che implementano le competenze specifiche, e la loro architettura di base è largamente determinata geneticamente con modulazione dall'esperienza durante periodi critici dello sviluppo.

Formalmente, la politica hardwired del Sistema M può essere rappresentata come una tabella di routing determinista:

$$\pi^M(s^M) = a^{M*} = \arg \max_{a^M \in A^M} Q^M(s^M, a^M)$$

dove $Q^M(s^M, a^M)$ è una funzione di valore delle meta-azioni che è fissa durante il ciclo di vita ma ottimizzata dall'evoluzione. Questa funzione Q^M non viene appresa per rinforzo durante la vita dell'agente: viene inizializzata dall'evoluzione (il parametro φ del framework bi-livello) e rimane costante. In alternativa, la politica può essere rappresentata come un insieme di regole condizionali della forma:

if $s^M \in R_k$ then $a^M = a^M_k$

dove $\{R_k\}$ è una partizione dello spazio dei meta-stati in regioni e $\{a^M_k\}$ è l'insieme delle meta-azioni associate. Questa rappresentazione è analoga alla tabella di routing in una SDN, e ha il vantaggio di essere interpretabile: ogni regola può essere mappata a un principio cognitivo biologicamente documentato.

Per esempio, la Tabella C.1 del paper suggerisce le seguenti regole di meta-controllo per i meta-stati biologici:

Per il meta-stato **sonno** (s^M : $t_{\text{active}} > \theta_{\text{sleep}}$): $a^M = (\text{close}(\text{Sensor} \rightarrow A), \text{close}(\text{Motor}), \text{open}(\text{Mem} \rightarrow A), \text{open}(A \rightarrow \text{Mem}), \text{mode}_A = \text{learn}, \text{mode}_B = \text{plan})$

Questa meta-azione chiude gli ingressi sensoriali e le uscite motorie, apre i canali tra la memoria episodica e il Sistema A, commuta il Sistema A in modalità di apprendimento e il Sistema B in modalità di pianificazione — corrispondente al replay delle esperienze recenti durante il sonno non-REM e alla pianificazione prospettica durante il sonno REM documentati da Wilson e McNaughton (1994) e da Pfeiffer e Foster (2013).

Per il meta-stato **dolore** (s^M : $F_{\text{contact}} > \theta_{\text{pain}}$): $a^M = (\text{interrupt}(B, \text{goal_directed}), \text{mode}_B = \text{exploit}, r^{\text{int_weight}} = 0, r_{\text{pain_boost}} = \infty, \text{write}(\text{Mem}, \text{current_experience}, \text{priority} = \text{max}))$

Questa meta-azione interrompe il comportamento goal-directed corrente, commuta il Sistema B in modalità di sfruttamento reattivo (la risposta di fuga), azzerà il reward intrinseco e assegna un segnale di pain come reward estrinseco negativo dominante, e scrive l'esperienza corrente nella memoria episodica con priorità massima — implementando il one-shot learning dal dolore documentato biologicamente da Eccleston e Crombez (1999).

Per il meta-stato **curiosità** (s^M : $\delta_A > \theta_{\text{novelty}}, \sigma^2_B < \theta_{\text{stable}}$): $a^M = (\text{mode}_B = \text{explore}, r^{\text{int_weight}} \rightarrow \text{max}, \lambda_A: \text{boost prediction loss weight})$

Questa meta-azione attiva la modalità esplorativa del Sistema B quando l'errore di previsione del Sistema A è elevato (novità) e il Sistema B è sufficientemente stabile da permettere l'esplorazione senza rischio, amplificando il reward intrinseco come segnale di curiosità.

7.5 Il problema dell'ottimizzazione bi-livello: scalabilità e stabilità

Il paper di Dupoux et al. propone, nella sezione 4, che il Sistema M venga ottimizzato alla scala evolutiva attraverso un'ottimizzazione bi-livello formalmente equivalente alla meta-learning. La struttura del problema è:

$$\varphi^* = \arg \min_{\varphi} L(A_0:A_K, B_0:B_K; \varphi)$$

soggetto a:

$$A_0, B_0, M = \text{Init}(\varphi)$$

$$A_{i+1}, B_{i+1} = \text{Update}(M, A_i, B_i, \text{Env}) \text{ per } i = 0, \dots, K-1$$

dove L è la fitness function calcolata sull'intero ciclo di vita dell'agente, K è il numero di step di aggiornamento del ciclo di vita, Env è l'ambiente con cui l'agente interagisce, e φ sono i parametri evolutivi che codificano la struttura del Sistema M , le condizioni iniziali dei sistemi A e B , e i parametri della fitness function.

Questo problema è un'istanza di ottimizzazione bi-livello (Sinha et al., 2018), in cui il livello esterno (evolutivo) minimizza la fitness function rispetto ai parametri φ , e il livello interno (sviluppativo) esegue l'aggiornamento dei sistemi A e B sotto il controllo di M . I metodi per risolvere problemi di questo tipo includono approcci gradient-based (MAML di Finn et al. (2017), iMAML di Rajeswaran et al. (2019)), approcci evolutivi (Neuroevolution, Stanley e Miikkulainen (2002)), e approcci ibridi (Evo-NAS, Liu et al. (2018)).

Le sfide computazionali di questo problema sono formidabili per almeno tre ragioni che il paper menziona senza sviluppare pienamente. La prima è la **non-differenziabilità del ciclo di vita**: la funzione $L(A_0:A_K, B_0:B_K; \varphi)$ non è differenziabile rispetto a φ attraverso il ciclo di vita interno a meno che l'Update sia differenziabile — il che è vero per gli aggiornamenti di discesa del gradiente del Sistema A (backpropagation), ma non necessariamente vero per gli aggiornamenti del Sistema B (reinforcement learning) né per le meta-azioni del Sistema M (routing discreto). I metodi gradient-based come MAML sono applicabili soltanto se il ciclo interno è differenziabile end-to-end, il che richiede che il Sistema M produca meta-azioni differenziabili — una condizione non soddisfatta dalla politica hardwired con routing discreto proposta nella sezione precedente.

I metodi evolutivi gradient-free — come CMA-ES (Covariance Matrix Adaptation Evolution Strategy, Hansen, 2006) o gli algoritmi genetici — non richiedono differenziabilità e sono quindi più adatti alla struttura del problema proposto. Tuttavia, la loro complessità campionaria cresce con la dimensionalità di φ : CMA-ES ha complessità $O(d^2)$ per step di ottimizzazione in dimensione d , il che è proibitivo per φ di grandi dimensioni. La riduzione della dimensionalità di φ — possibile attraverso l'imposizione di struttura sulle politiche del Sistema M , come la rappresentazione a tabella di routing proposta nella sezione precedente — è essenziale per la trattabilità computazionale del problema evolutivo.

La seconda sfida è la **non-stazionarietà dell'ambiente di fitness**: la funzione L dipende dall'Env, e se l'ambiente varia nel corso dell'ottimizzazione evolutiva — come proposto dal meccanismo di

Evolutionary Curriculum del paper — allora il problema di ottimizzazione è non-stazionario: il gradiente (o il paesaggio di fitness) cambia nel corso dell'ottimizzazione. Questo rende difficile la convergenza degli algoritmi standard, e richiede metodi appositamente progettati per problemi di ottimizzazione curriculare — come Open-Ended Learning (OEL) di Wang et al. (2019) e gli algoritmi PAIRED di Dennis et al. (2020).

La terza sfida è la **lunghezza del ciclo di vita interno**: il ciclo interno deve essere sufficientemente lungo (K sufficientemente grande) da permettere ai sistemi A e B di acquisire competenze significative, ma sufficientemente breve da rendere computazionalmente trattabile l'ottimizzazione del livello esterno. Il numero di cicli di vita necessari per l'ottimizzazione del livello esterno cresce con la varianza della fitness — quanto più variabile è la fitness attraverso i cicli di vita, tanto più campioni sono necessari per stimare in modo affidabile il gradiente del livello esterno. La riduzione della varianza della fitness — possibile attraverso il controllo accurato degli ambienti di training, la standardizzazione della valutazione e l'uso di baseline computazionali per la riduzione della varianza — è un prerequisito pratico per la scalabilità del training evolutivo.

7.6 Il Sistema M e la prefrontal cortex: neurobiologia del meta-controllo

Il paper di Dupoux et al. stabilisce un'analogia esplicita tra il Sistema M e la corteccia prefrontale (PFC), citando la revisione di Miller e Cohen (2001). Questa analogia merita di essere esaminata in dettaglio, sia per le sue virtù esplicative sia per i suoi limiti.

La corteccia prefrontale — in particolare le regioni dorso-laterale (DLPFC), orbito-frontale (OFC) e anteriore cingolata (ACC) — svolge nei primati una gamma di funzioni esecutive che includono: la pianificazione e la selezione delle azioni in funzione di obiettivi a lungo termine (DLPFC), la valutazione affettiva delle azioni e la modulazione delle risposte in funzione del loro valore (OFC), e il monitoraggio dei conflitti tra risposte competitive e la modulazione dell'attenzione in funzione di questo monitoraggio (ACC). Queste funzioni corrispondono grosso modo alle funzioni del Sistema M: il routing delle risorse cognitive in funzione dei meta-stati, la modulazione delle funzioni di perdita in funzione del valore delle meta-azioni, e il monitoraggio del conflitto tra i sistemi A e B come segnale di meta-stato epistemico.

Il modello di Miller e Cohen (2001) — che formalizza il ruolo della PFC come "bias signals" che modulano l'attività di aree corticali sensibili-al-compito per mantenere la pertinenza-al-compito delle elaborazioni in corso — è strutturalmente analogo al meccanismo di routing del Sistema M: il Sistema M genera segnali di meta-azione che modulano il flusso di informazioni tra i componenti, favorendo i percorsi di elaborazione pertinenti all'obiettivo corrente e sopprimendo i percorsi competitivi. La

differenza chiave è che nel modello biologico i bias signals della PFC sono continui e operano attraverso la modulazione neuromodulatoriale (dopaminergica, noradrenergica, serotoninergica) delle reti corticali, mentre il Sistema M proposto nel paper genera meta-azioni discrete che modificano la topologia connessionale del sistema.

Il meccanismo neuromodulatoriale è particolarmente rilevante perché offre una via di implementazione per le meta-azioni di modulazione della perdita (A^M_{loss}) che supera i limiti delle azioni discrete. Nel cervello biologico, la dopamina modula la plasticità sinaptica nelle proiezioni corticostriatali, implementando di fatto una modulazione del tasso di apprendimento e del peso relativo degli obiettivi di rinforzo. La noradrenalina modula il rapporto segnale-rumore nelle reti corticali, implementando una forma di controllo del "guadagno" che ha effetti analoghi alla modulazione della temperatura dei sampling distributions nei modelli di apprendimento per rinforzo. La serotonina modula la tolleranza al rischio e l'orizzonte temporale dell'ottimizzazione, implementando una forma di controllo del discount factor γ .

Un'architettura computazionale che mimasse questa struttura neuromodulatoriale potrebbe implementare le meta-azioni di modulazione della perdita attraverso segnali continui che modulano i parametri della backpropagation e del reinforcement learning piuttosto che attraverso azioni discrete: il Sistema M genererebbe vettori di modulazione continui (analogici ai livelli di neurotrasmettitori) che scalano i gradienti del Sistema A e del Sistema B in funzione del meta-stato corrente. Questa architettura sarebbe differenziabile end-to-end — consentendo l'ottimizzazione bi-livello gradient-based — a differenza dell'architettura a routing discreto.

7.7 Modalità operative avanzate: apprendimento per comunicazione e per immaginazione come regimi di routing

L'Appendice B del paper di Dupoux et al. descrive due modalità operative avanzate del Sistema M — l'apprendimento per comunicazione e l'apprendimento per immaginazione — affermando che queste modalità emergono dalla combinazione delle funzioni di base del Sistema M. La formalizzazione di queste modalità in termini dello spazio delle meta-azioni definito nelle sezioni precedenti rivela la struttura computazionale che le sottende.

L'apprendimento per comunicazione corrisponde al regime di routing in cui i segnali pedagogici del conspecifico-docente (gaze, pointing, tono imperativo, contesto di scaffolding) attivano il meta-stato $\sigma^{\sigma_teacher}$, che a sua volta genera la meta-azione:

$$a^M_{communication} = ($$

```

open(Teacher_signal → A, priority_boost = trust(Teacher)),
mode_A = learn,
λ_A: boost weight(Teacher_signal loss),
write(Mem, Teacher_signal, priority ∝ trust(Teacher))
)

```

Il parametro $\text{trust}(\text{Teacher})$ — la stima della affidabilità del docente — è essa stessa un meta-stato che evolve nel tempo in funzione della congruenza tra le predizioni del docente e le osservazioni successive. Questo implementa il meccanismo di epistemic vigilance documentato da Sperber et al. (2010): il Sistema M non tratta tutte le sorgenti di segnale pedagogico come ugualmente credibili, ma pesa i segnali in funzione della stima della affidabilità della sorgente.

La formalizzazione del $\text{trust}(\text{Teacher})$ richiede un meccanismo di aggiornamento bayesiano. Si denoti con H^t il vettore delle ipotesi del sistema sulle prossime osservazioni condizionate sull'istruzione del docente D^t : $H^t = P(S_{t+1} \mid D^t, s^M_t)$. Si denoti con O_{t+1} l'osservazione effettiva. Il trust viene aggiornato:

$$\text{trust}_{t+1}(\text{Teacher}) \propto \text{trust}_t(\text{Teacher}) \cdot \frac{P(O_{t+1} \mid H^t)}{P(O_{t+1})}$$

ovvero il trust viene amplificato quando l'istruzione del docente genera previsioni che si rivelano accurate, e ridotto quando le previsioni sono inaccurate. Questo meccanismo implementa l'apprendimento selettivo dai conspecifici affidabili documentato da Koenig et al. (2004) e Zmyj et al. (2010).

L'apprendimento per immaginazione corrisponde al regime di routing in cui i canali sensoriali sono chiusi o attenuati e il Sistema A e il Sistema B operano su input provenienti dalla memoria episodica piuttosto che dai sensori:

```

a^M_imagination = (
close(Sensor → A) ∨ attenuate(Sensor → A, α_att < 1),
open(Mem → A),
open(A_latent → B_input),
close(B_output → Motor) ∨ redirect(B_output → A_input),
mode_A = learn + infer,

```

mode_B = plan

)

In questa configurazione, il Sistema A genera rappresentazioni latenti da memorie episodiche recuperate, il Sistema B elabora queste rappresentazioni come se fossero osservazioni correnti e genera piani di azione, e le azioni pianificate vengono reindirizzate come input aggiuntivi al Sistema A per una valutazione predittiva della loro efficacia — invece di essere inviate ai motori. Il risultato è un ciclo di simulazione interna in cui il sistema valuta le conseguenze predette delle azioni pianificate senza eseguirle nel mondo reale — l'analogo computazionale della pianificazione mentale e del daydreaming.

Questa configurazione corrisponde, nella neuroscienza cognitiva, all'attività della Default Mode Network (DMN) durante il mind-wandering e la simulazione prospettica documentata da Buckner et al. (2008) e Raichle (2015): la DMN è particolarmente attiva durante i periodi di riposo quando non è richiesta elaborazione sensoriale esterna, e la sua attività è correlata con il replay di eventi passati e la simulazione di scenari futuri. Il Sistema M implementa questa configurazione attraverso il meta-stato somatico di riposo ($s^{\omega}: t_{\text{active}} < \theta_{\text{rest}}, F_{\text{contact}} \approx 0$) che attiva il routing verso la modalità di immaginazione.

7.8 Il Sistema M come soggetto di apprendimento continuo: la questione della meta-plasticità

Una questione che il paper di Dupoux et al. affronta soltanto tangenzialmente — e che rappresenta uno dei nodi teorici più profondi del framework A-B-M — è se il Sistema M possa o debba essere soggetto a forme limitate di plasticità durante il ciclo di vita, nonostante la politica hardwired proposta nella sezione 15.4. Il paper accenna in una nota a piè di pagina alla possibilità che l'acquisizione di strategie di apprendimento specifiche della cultura — come l'andare a scuola — sia compatibile con un Sistema M fisso, senza tuttavia sviluppare questa affermazione in modo formalmente rigoroso.

La questione può essere inquadrata attraverso il concetto di **meta-plasticità** introdotto da Abraham e Bear (1996) in neurobiologia: la plasticità sinaptica — la capacità di modificare l'efficacia di una connessione sinaptica in risposta all'attività — è essa stessa modulata dall'attività precedente, in modo tale che l'istoria dell'attività di un neurone influenza la sua futura capacità di plasticità. In termini computazionali, la meta-plasticità implica che il tasso di apprendimento di un sistema non è un parametro fisso ma è esso stesso modulato dall'esperienza — una forma di plasticità del secondo ordine.

Applicata al Sistema M, la meta-plasticità suggerirebbe che alcune componenti della politica del Sistema M — in particolare quelle relative alla

modulazione del trust e alla calibrazione dei segnali epistemici — potrebbero essere soggette a aggiornamenti durante il ciclo di vita, mentre le componenti di routing fondamentali (la risposta ai segnali di dolore, la commutazione in modalità di sonno, la risposta a stimoli di minaccia) rimangono fisse. Questa distinzione corrisponde alla differenza biologica tra la plasticità dei meccanismi di controllo esecutivo PFC-dipendente — che è relativamente alta e si modifica con l'esperienza culturale — e la plasticità dei meccanismi limbici e sottocorticali che implementano le risposte istintive — che è bassa e principalmente determinata geneticamente.

Formalmente, questa distinzione può essere implementata partizionando la politica del Sistema M in due componenti:

$$\pi^M = \pi^M_{\text{fixed}} \oplus \pi^M_{\text{adaptive}}$$

dove π^M_{fixed} è la componente hardwired ottimizzata evolutivamente e invariante durante il ciclo di vita, e π^M_{adaptive} è una componente adattiva che può essere aggiornata durante il ciclo di vita attraverso un meccanismo di meta-apprendimento di terzo livello. Il meccanismo di aggiornamento di π^M_{adaptive} potrebbe essere un algoritmo di meta-learning come MAML o Reptile, addestrato non sulla fitness function del ciclo di vita ma su un obiettivo di meta-meta-controllo — ad esempio, la capacità di calibrare rapidamente le stime di trust in nuovi contesti sociali.

Questa struttura gerarchica — politica di meta-controllo fissa + componente adattiva + meta-meta-apprendimento della componente adattiva — introduce una nuova regressione, ma una regressione finita e controllata: a differenza della regressione infinita che il paper mira ad evitare con la politica hardwired, questa è una regressione a tre livelli in cui il livello più alto (la struttura di π^M_{adaptive}) è determinata evolutivamente e rimane fissa, soltanto i parametri di π^M_{adaptive} vengono aggiornati durante il ciclo di vita, e i parametri del meccanismo di aggiornamento sono parte del parametro φ ottimizzato alla scala evolutiva.

7.9 Confronto con architetture cognitive classiche: ACT-R, SOAR e Global Workspace Theory

Il Sistema M di Dupoux et al. non emerge nel vuoto: si inserisce in una tradizione di architetture cognitive computazionali che hanno tentato di formalizzare i meccanismi di controllo esecutivo e di coordinazione dei processi cognitivi. Il confronto con le architetture più rilevanti — ACT-R di Anderson, SOAR di Laird, e la Global Workspace Theory di Baars e Dehaene — illumina sia i contributi originali del framework A-B-M sia le sue continuità con le tradizioni precedenti.

L'architettura **ACT-R** (Adaptive Control of Thought-Rational, Anderson et al., 2004) è un'architettura cognitiva modulare in cui processi di conoscenza

specializzati (percezione, motricità, memoria dichiarativa, memoria procedurale) interagiscono attraverso un buffer centrale che serializza le richieste di accesso alle memorie e coordina l'esecuzione delle procedure. Il meccanismo di controllo di ACT-R è basato su regole di produzione (production rules) — coppie condizione-azione della forma "if X then do Y" — che specificano come il sistema deve rispondere alle configurazioni dei buffer centrali. La selezione delle regole di produzione è probabilistica e dipende dall'utilità attesa di ciascuna regola — una quantità che viene appresa per rinforzo attraverso l'interazione con il sistema.

Il Sistema M del framework A-B-M è strutturalmente analogo al meccanismo di produzione di ACT-R: le regole di routing del Sistema M sono coppie meta-stato/meta-azione della forma " $s^M \in R_k$ then $a^M = a^M_k$ ", equivalenti alle production rules di ACT-R. La differenza principale è che in ACT-R le regole di produzione vengono apprese per rinforzo durante il ciclo di vita del sistema (all'interno del sistema), mentre nel framework A-B-M le regole di routing del Sistema M sono ottimizzate evolutivamente (al di fuori del ciclo di vita). Questa differenza riflette la diversa posizione epistemologica delle due architetture: ACT-R è progettata per modellare l'apprendimento cognitivo umano nell'arco di un'esperienza di vita, mentre il framework A-B-M è progettato per costruire agenti che possano apprendere autonomamente in modo analogo agli organismi biologici.

L'architettura **SOAR** (State, Operator And Result, Laird, 2012) è un'architettura a spazio dei problemi in cui il controllo viene esercitato attraverso la selezione di operatori che trasformano gli stati del problema. La caratteristica distintiva di SOAR è il meccanismo di **chunking**: quando il sistema risolve un sotto-problema attraverso un processo deliberativo (impasse resolution), il risultato di questa deliberazione viene compilato in un chunk — una regola di produzione che può essere applicata direttamente in futuro senza deliberazione esplicita. Il chunking implementa una forma di compilazione della conoscenza dall'esplicito all'implicito — dalla deliberazione alla procedura automatica — che è analogamente documentata nei processi di skill acquisition umani.

Il meccanismo di chunking di SOAR corrisponde, nel framework A-B-M, al trasferimento di conoscenza dal Sistema A (deliberazione, pianificazione esplicita) al Sistema B (politiche reattive). Quando il Sistema A ha prodotto, attraverso la simulazione, un piano che si è rivelato efficace, questa conoscenza pianificatoria può essere — in linea di principio — distillata in una politica del Sistema B che produca lo stesso comportamento direttamente, senza richiedere la simulazione esplicita del Sistema A. Questo tipo di distillazione non è discusso esplicitamente nel paper di Dupoux et al., ma è implicito nella struttura dell'interazione tra i sistemi e rappresenta uno dei meccanismi attraverso cui il framework A-B-M potrebbe realizzare

l'analogia biologica tra pianificazione cosciente e comportamento automatizzato.

La **Global Workspace Theory** (GWT, Baars, 1988; Dehaene e Changeux, 2011) propone che la coscienza emerga da un "workspace globale" — un sistema di broadcasting che distribuisce informazione in modo ampio e non-specifico attraverso l'architettura modulare del cervello, consentendo la coordinazione flessibile tra processi locali specializzati che altrimenti opererebbero in modo indipendente. Il workspace globale corrisponde, nella terminologia di Dehaene, ai neuroni dello strato 2/3 delle aree frontali e parietali che proiettano ampiamente attraverso la corteccia, implementando una forma di attenzione globale sulle rappresentazioni locali dei moduli specializzati.

Il Sistema M del framework A-B-M è strutturalmente analogo al workspace globale della GWT: svolge la funzione di broadcasting coordinato tra i componenti del sistema — apre canali che distribuiscono le rappresentazioni del Sistema A al Sistema B e viceversa, e distribuisce i segnali di meta-controllo a entrambi i sistemi. Il paper cita esplicitamente la Global Workspace Theory architecture di Goyal e Bengio (2022) come un tentativo di implementare questo tipo di coordinazione nel contesto del deep learning, notando che questa linea di ricerca è ancora in fase embrionale di sviluppo. La connessione tra il Sistema M e la GWT suggerisce che il framework A-B-M potrebbe beneficiare dell'incorporazione di meccanismi di broadcasting globale — come i meccanismi di attenzione multi-testa che il Transformer implementa a livello linguistico, ma applicati al coordinamento dell'intera architettura piuttosto che alle relazioni intra-sequenza.

La formalizzazione del Sistema M che questa sezione ha sviluppato offre un punto di partenza per integrare queste tradizioni in un framework unificato: le regole di routing del Sistema M corrispondono alle production rules di ACT-R e agli operatori di SOAR; il meccanismo di distillazione implicito nel trasferimento da Sistema A a Sistema B corrisponde al chunking di SOAR; e il broadcasting globale del Sistema M corrisponde al workspace globale della GWT. Questa convergenza suggerisce che le diverse architetture cognitive, sviluppate in tradizioni di ricerca indipendenti, stiano convergendo verso una struttura comune che il framework A-B-M articola con un livello di specificità formale superiore a quello delle architetture precedenti, pur lasciando aperte — onestamente — le questioni di implementazione più difficili.

Sezione VIII — L'Integrazione LLM-JEPA: Ibridi Architetture, Convergenza dei Paradigmi e il Programma di Ricerca della Sintesi

8.1 Il problema della complementarità strutturale

I capitoli precedenti hanno documentato con sistematicità le virtù e i limiti di ciascuno dei due paradigmi esaminati — i Large Language Models e le architetture World Model con JEPA nel framework A-B-M — mostrando che le debolezze dell'uno corrispondono spesso, con notevole precisione, alle forze dell'altro. Gli LLM eccellono nell'elaborazione del linguaggio naturale, nella sintesi di conoscenza distribuita su scala Internet, nella generalizzazione zero-shot a task linguistici nuovi e nella produzione di ragionamento esplicito multi-passo attraverso il chain-of-thought; ma sono strutturalmente privi di grounding fisico, incapaci di apprendimento continuo post-dispiegamento, fragili rispetto a distributional shift in domini non linguistici, e confinati al primo livello della gerarchia causale di Pearl nella misura in cui la loro conoscenza causale è costruita esclusivamente dalla statistica distribuzionale del corpus testuale. I World Models con JEPA nel framework A-B-M sono progettati per superare esattamente questi limiti: costruire rappresentazioni ancorate all'esperienza sensomotoria, apprendere in modo continuo dall'interazione con l'ambiente, generalizzare in modo robusto attraverso la struttura invariante fisica del mondo, e accedere — attraverso l'intervento del Sistema B — al secondo livello della gerarchia causale. Ma al momento del dispiegamento pratico, mancano della ricchezza semantica e della competenza linguistica degli LLM.

La domanda naturale che emerge da questa analisi comparativa è se sia possibile — e in quale forma — costruire sistemi che combinino i vantaggi di entrambi i paradigmi senza ereditarne i limiti. Questa domanda non è nuova: l'integrazione tra ragionamento simbolico-linguistico e percezione grounded era già l'ambizione dei sistemi neuro-simbolici degli anni Novanta, e il fallimento di quell'integrazione ha contribuito all'ascesa del paradigma puramente connessionista. Tuttavia, la domanda si pone oggi in condizioni radicalmente diverse: i due paradigmi da integrare sono entrambi connessionisti, entrambi basati su trasformatori e meccanismi di attenzione, e entrambi addestrati su scala senza precedenti — il che elimina molti degli ostacoli tecnici che rendevano difficile l'integrazione in epoche precedenti.

8.2 Tassonomia degli approcci di integrazione LLM-JEPA

Prima di analizzare le architetture di integrazione specifiche, è utile costruire una tassonomia degli approcci possibili lungo due dimensioni: il **grado di accoppiamento** tra i due sistemi (da accoppiamento lasco a integrazione

profonda) e la **direzione del flusso informativo dominante** (dall'LLM al World Model, dal World Model all'LLM, o bidirezionale).

Il **accoppiamento lasco** corrisponde ad architetture in cui l'LLM e il World Model operano come sistemi separati che si scambiano informazioni attraverso interfacce ben definite, senza condivisione di rappresentazioni interne. Un esempio paradigmatico è l'architettura Vision-Language-Action (VLA) del tipo RT-2 di Zitkovich et al. (2023): il modello linguistico elabora le istruzioni linguistiche e le immagini visive, e le sue rappresentazioni interne vengono usate per condizionare la generazione di azioni motorie attraverso un policy network separato. In questa architettura, l'LLM non modifica il suo modello del mondo in risposta all'esperienza motoria, e il policy network non accede alle rappresentazioni linguistiche interne dell'LLM se non attraverso le azioni che l'LLM genera. L'accoppiamento è lasco perché i due sistemi mantengono architetture interne separate e non si influenzano vicendevolmente durante l'inferenza.

L'**integrazione profonda** corrisponde invece ad architetture in cui le rappresentazioni latenti dell'LLM e del World Model sono congiuntamente ottimizzate e condivise, in modo che ciascun sistema influenzi la struttura delle rappresentazioni dell'altro. Un esempio concettuale — non ancora pienamente realizzato nella letteratura corrente — sarebbe un'architettura in cui l'encoder del World Model JEPA e l'encoder dell'LLM condividono gli stessi parametri o vengono addestrati congiuntamente su un obiettivo combinato che include sia la previsione linguistica sia la previsione di stato fisico futuro.

Lungo la dimensione della direzione del flusso, si distinguono tre configurazioni fondamentali:

LLM → World Model: l'LLM fornisce prior semantici, obiettivi linguisticamente specificati, e conoscenza di commonsense che guidano l'apprendimento e la pianificazione del World Model. In questa configurazione, l'LLM svolge il ruolo di "configuratore" nel senso di LeCun (2022): traduce obiettivi linguisticamente specificati in target per il Sistema B del World Model.

World Model → LLM: il World Model fornisce all'LLM rappresentazioni grounded del mondo fisico che arricchiscono il ragionamento linguistico con struttura sensomotoria. In questa configurazione, il World Model funge da "grounding engine" per l'LLM: le sue rappresentazioni latenti degli oggetti, delle loro proprietà fisiche e delle relazioni causali tra stati e azioni vengono usate per condizionare la generazione linguistica dell'LLM verso descrizioni più accurate e fisicamente coerenti.

Bidirezionale: il flusso informativo è simmetrico, con aggiornamenti reciproci tra le rappresentazioni latenti dei due sistemi durante sia l'addestramento che l'inferenza.

8.3 L'approccio "LLM come configuratore del World Model": formalizzazione e implementazioni

La proposta più articolata di integrazione unidirezionale $LLM \rightarrow World Model$ è quella sviluppata da LeCun (2022) nel documento "A Path Towards Autonomous Machine Intelligence": l'LLM svolge il ruolo di **configuratore** che traduce obiettivi linguisticamente specificati in obiettivi nello spazio latente del World Model, permettendo al Sistema B di ottimizzare politiche per obiettivi specificati in linguaggio naturale senza la necessità di progettare manualmente funzioni di reward per ogni compito.

Formalmente, il configuratore è una funzione $C_\omega : L \times Z \rightarrow Z$ che mappa una descrizione linguistica del goal $g \in L$ e uno stato corrente $z_t \in Z$ in uno stato target $\hat{z}_g \in Z$ nello spazio latente del World Model:

$$\hat{z}_g = C_\omega(g, z_t)$$

Il Sistema B può quindi ottimizzare la policy per raggiungere $\hat{z}_g$ minimizzando la distanza nel spazio latente:

$$\pi^*(a \mid z_t) = \arg \min_{\pi} E_{\pi} [d(z_T, \hat{z}_g)]$$

dove z_T è lo stato latente raggiunto al termine della traiettoria e $d(\cdot, \cdot)$ è una misura di distanza nello spazio latente del World Model (tipicamente la distanza coseno o la distanza euclidea normalizzata per le rappresentazioni JEPA).

La funzione C_ω può essere implementata come una rete neurale che mappa embeddings linguistici dell'LLM in vettori nello spazio latente del World Model. L'addestramento di C_ω richiede coppie $(g, \hat{z}_g)$ dove g è una descrizione linguistica di uno stato target e $\hat{z}_g$ è la corrispondente rappresentazione latente nel World Model — dati che possono essere generati attraverso una combinazione di dati di teleoperation etichettati (dove un operatore umano descrive verbalmente l'obiettivo che sta perseguendo) e di auto-annotazione (dove l'LLM genera automaticamente descrizioni degli stati visitati durante l'esplorazione autonoma del Sistema B).

L'implementazione più avanzata di questo approccio disponibile nella letteratura corrente è SayCan di Ahn et al. (2022), in cui un LLM (PaLM) genera piani linguistici per compiti di manipolazione robotica in cucina, e affordance functions pre-addestrate valutano la fattibilità di ciascuna azione linguistica nell'ambiente fisico corrente, permettendo di selezionare il piano che massimizza sia la probabilità linguistica (probabilità che l'LLM assegna

all'azione data l'istruzione) sia la probabilità di successo fisico (probabilità che l'affordance function assegna all'azione nello stato corrente):

$$a^* = \arg \max_a [P_{\text{LLM}}(a \mid \text{instruction}) \cdot P_{\text{aff}}(\text{success} \mid a, s)]$$

SayCan non usa un World Model JEPA — usa affordance functions specifiche per ogni azione, pre-addestrate separatamente — ma la struttura del problema che risolve è esattamente quella del configuratore di LeCun: usare la conoscenza semantica dell'LLM per guidare la selezione di azioni fisicamente fattibili nell'ambiente corrente.

8.4 L'approccio "World Model come grounding engine per l'LLM": rappresentazioni multimodali e fisica intuitiva

La direzione opposta di integrazione — World Model $\rightarrow$ LLM — parte dall'osservazione che le rappresentazioni interne degli LLM, costruite esclusivamente su statistiche distribuzionali del testo, difettano di struttura fisica coerente: le relazioni spaziali, causali e temporali tra oggetti fisici che un bambino di tre anni comprende attraverso l'interazione diretta con il mondo richiedono, per un LLM, enormi quantità di descrizioni testuali — e rimangono comunque fragili rispetto a variazioni percettive che il bambino gestirebbe senza difficoltà.

L'idea fondamentale di questo approccio è di usare le rappresentazioni latenti di un World Model pre-addestrato — che ha acquisito struttura fisica attraverso l'apprendimento di previsioni su dati sensomotori — come **rappresentazioni grounded** che vengono iniettate nell'architettura dell'LLM per arricchire la sua comprensione di riferimenti fisici nel testo. Un esempio paradigmatico è l'uso di rappresentazioni visive pre-addestrate come input aggiuntivi per modelli linguistici — nella tradizione di CLIP (Radford et al., 2021), Flamingo (Alayrac et al., 2022) e LLaVA (Liu et al., 2023). Tuttavia, questi sistemi usano encoder visivi pre-addestrati su obiettivi di classificazione o contrastivi, non World Model nel senso di JEPA: le loro rappresentazioni visive catturano categorie semantiche ma non la struttura dinamica e causale del mondo fisico.

Un'integrazione più profonda userebbe le rappresentazioni di un World Model JEPA addestrato su traiettorie sensomotorie — rappresentazioni che codificano le proprietà fisiche degli oggetti, le loro affordance e le regolarità causali delle loro interazioni — come prior fisica per il ragionamento dell'LLM. Formalmente, questo può essere implementato come segue. Sia $z_V = g_\varphi(v)$ la rappresentazione latente di un input visivo v prodotta dall'encoder del World Model JEPA. Sia x_L la sequenza di token dell'LLM. Invece di condizionare l'LLM soltanto su x_L , si condiziona su (x_L, z_V) :

$$P_\theta(x_t \mid x_{<t}) \rightarrow P_{\{\theta, \psi\}}(x_t \mid x_{<t}, z_V)$$

dove ψ sono i parametri di un cross-attention module che proietta z_V nello spazio delle chiavi e dei valori del meccanismo di attenzione dell'LLM:

$$K_V = W_K \cdot z_V, V_V = W_V \cdot z_V$$

$$h'_t = \text{MultiHead}(Q_t, [K_{\text{text}}, K_V], [V_{\text{text}}, V_V])$$

Questo meccanismo — equivalente all'architettura cross-attention di Flamingo ma con z_V prodotto da un World Model JEPA invece che da un encoder visivo classificativo — permetterebbe all'LLM di condizionare il proprio ragionamento linguistico su rappresentazioni della struttura fisica del mondo acquisite attraverso l'apprendimento sensomotorio, non soltanto attraverso la statistica del testo.

Il vantaggio atteso di questa integrazione rispetto all'uso di encoder visivi classificativi è la maggiore ricchezza della struttura fisica nelle rappresentazioni JEPA: poiché il World Model JEPA è addestrato a prevedere stati fisici futuri condizionatamente alle azioni, le sue rappresentazioni codificano le proprietà **dinamiche** degli oggetti — le traiettorie tipiche, le forze di contatto, le deformazioni — piuttosto che soltanto le proprietà **statiche** (categoria, colore, forma) che un encoder classificativo apprende. Questa struttura dinamica è esattamente ciò che manca alle descrizioni testuali su cui gli LLM vengono addestrati: è difficile descrivere in testo il preciso peso di un oggetto, la sua deformabilità, il suo coefficiente di attrito — ma queste proprietà sono acquisite implicitamente da un World Model che ha osservato migliaia di traiettorie di manipolazione dell'oggetto.

8.5 Transfusion: verso una previsione congiunta testo-immagine

Un sviluppo recente che rappresenta un passo concreto verso l'integrazione profonda — sebbene non ancora nel framework JEPA — è l'architettura Transfusion di Zhou et al. (2024), citata nel paper di Dupoux et al. nella Tabella 1 come esempio di apprendimento multimodale. Transfusion addestra un singolo modello con un unico set di parametri su due obiettivi simultanei: previsione autoregressiva del prossimo token per il testo (l'obiettivo LLM standard) e previsione di rumore per la denoising diffusion per le immagini (l'obiettivo del modello generativo di immagini). La chiave è che i due obiettivi condividono la stessa architettura Transformer e gli stessi pesi — il che implica che le rappresentazioni interne del modello devono essere contemporaneamente utili per la previsione linguistica e per la generazione/comprendimento di immagini.

I risultati di Transfusion mostrano che questa co-ottimizzazione produce rappresentazioni che superiori a modelli specializzati su benchmark sia linguistici sia visivi, suggerendo che la condivisione dei parametri produce una forma di **mutual regularization**: l'obiettivo linguistico vincola le rappresentazioni verso strutture semantiche coerenti con il linguaggio,

mentre l'obiettivo visivo vincola le rappresentazioni verso strutture percettive fisicamente strutturate. Il risultato è una rappresentazione che cattura entrambi i tipi di struttura in modo più efficiente rispetto a sistemi separati.

L'estensione di questa logica al framework JEPA — addestrando un singolo modello con l'obiettivo autoregressivo testuale dell'LLM e l'obiettivo di previsione latente del JEPA — è concettualmente naturale ma tecnicamente non triviale. Il principale ostacolo è la differenza strutturale tra i due obiettivi: l'obiettivo LLM è una distribuzione di probabilità su token discreti (una distribuzione categoriale su $|V|$ classi), mentre l'obiettivo JEPA è una previsione deterministica (o stocastica) in uno spazio latente continuo. La co-ottimizzazione richiede che l'architettura produca sia rappresentazioni discrete-compatibili per l'obiettivo linguistico sia rappresentazioni continue-predittive per l'obiettivo JEPA — una dualità che architetture come Transfusion realizzano attraverso teste di decodifica separate ma rappresentazioni interne condivise.

La formula dell'obiettivo combinato sarebbe:

$$L_{\text{combined}} = \lambda_{\text{lang}} \cdot L_{\text{LM}} + \lambda_{\text{JEPA}} \cdot L_{\text{JEPA}} + \lambda_{\text{cross}} \cdot L_{\text{cross}}$$

dove L_{LM} è la perdita di modellazione del linguaggio, L_{JEPA} è la perdita di previsione latente del JEPA, e L_{cross} è un termine di perdita che penalizza la divergenza tra le previsioni del dominio linguistico e del dominio fisico quando entrambi si riferiscono allo stesso referente — implementando un vincolo di coerenza cross-modale. Il parametro λ_{cross} è critico: se è troppo basso, i due obiettivi sono indipendenti e non si influenzano reciprocamente; se è troppo alto, il modello può sacrificare la qualità di un obiettivo per soddisfare l'altro. La calibrazione di questi pesi è un problema di ottimizzazione multi-obiettivo che richiede metodi specifici come il Pareto front optimization o il gradiente multi-task.

8.6 L'architettura LLM-JEPA nel framework A-B-M: una proposta integrativa

Integrando le analisi delle sezioni precedenti, è possibile delineare un'architettura che realizzi la sintesi tra LLM e World Model JEPA nel contesto del framework A-B-M proposto da Dupoux et al. Questa architettura non è un sistema esistente ma un programma di ricerca: il tentativo di specificare formalmente come l'integrazione dovrebbe essere strutturata per realizzare le promesse teoriche della sintesi.

L'architettura proposta ha quattro componenti principali:

Componente 1 — Encoder Multimodale Condiviso (EMC): un encoder Transformer che elabora sia sequenze di token testuali sia flussi di frame visivi (o altri input sensoriali), producendo rappresentazioni latenti in uno

spazio comune Z_{shared} . Questo encoder è addestrato congiuntamente su un obiettivo LM per le sequenze testuali e un obiettivo JEPA per le sequenze visive, con l'obiettivo L_{cross} che impone coerenza cross-modale quando lo stesso concetto è presente in entrambe le modalità. L'EMC svolge il ruolo del Sistema A nel framework A-B-M, fornendo rappresentazioni del mondo sia per il ragionamento linguistico sia per la pianificazione motoria.

Componente 2 — Predittore Dinamico (PD): un Transformer condizionato sull'azione che, dato lo stato latente $z_t \in Z_{\text{shared}}$ e un'azione a_t (codificata come token discreto o come vettore continuo), prevede lo stato latente futuro $\hat{z}_{t+1} \in Z_{\text{shared}}$. Il PD apprende la dinamica del mondo nello spazio latente condiviso, e può essere usato dal Sistema B per il model-based planning attraverso il roll-out simulato di traiettorie di azioni.

Componente 3 — Modulo Linguistico Autoregressivo (MLA): un decoder autoregressivo che, dato un contesto di token testuali e le rappresentazioni latenti dell'EMC, genera il prossimo token testuale. Questo modulo implementa la funzione configuratore di LeCun: traduce obiettivi linguisticamente specificati in rappresentazioni nello spazio latente di Z_{shared} che il Sistema B può usare come target per l'ottimizzazione della policy. Il MLA può anche essere usato per generare spiegazioni testuali del processo di pianificazione del Sistema B — una funzione di explainability integrata nell'architettura.

Componente 4 — Sistema B Incarnato (SBI): un agente di reinforcement learning che opera su rappresentazioni di Z_{shared} prodotte dall'EMC, usa il PD per model-based planning, e riceve obiettivi da Z_{shared} prodotti dalla traduzione del MLA delle istruzioni linguistiche. Il SBI implementa il Sistema B del framework A-B-M con l'architettura di policy optimization di scelta (PPO, SAC, o DDPG per spazi di azione continui).

L'orchestrazione dei quattro componenti è affidata al Sistema M, che monitora i meta-stati epistemici (errori di previsione dell'EMC e del PD, qualità del reward del SBI, coerenza cross-modale tra le rappresentazioni testuali e visive dell'EMC), somatici (stato fisico del robot) e specifici della specie (segnali pedagogici da un eventuale docente umano), e genera meta-azioni che configurano il routing tra i componenti e modulano gli obiettivi di addestramento.

8.7 RT-2 e SayCan come stepping stones verso l'integrazione profonda

I sistemi RT-2 (Zitkovich et al., 2023) e SayCan (Ahn et al., 2022) rappresentano, nella letteratura corrente, i tentativi più maturi di integrazione tra linguaggio e azione — pur non realizzando ancora l'integrazione profonda dell'architettura LLM-JEPA delineata nella sezione precedente. La

loro analisi in dettaglio permette di identificare esattamente le componenti mancanti per il passo successivo.

RT-2 fine-tunes un Vision-Language Model (VLM) pre-addestrato — specificamente PaLM-E e Gemini Visual — su dati di robotica in modo che il modello produca direttamente token di azione (discretizzati in 256 classi per ciascuna dimensione del vettore di azione 7-DOF) come se fossero token linguistici. L'architettura è elegante nella sua semplicità: non richiede componenti aggiuntivi, soltanto un fine-tuning dell'obiettivo di modellazione del linguaggio per includere token di azione nel vocabolario. I risultati sono notevoli: RT-2 generalizza a task descritti linguisticamente ma mai visti durante il training robotico — ad esempio, eseguire compiti di manipolazione che richiedono conoscenza implicita di commonsense ("porta l'oggetto che può essere usato come cibo per un uccello nel contenitore") — con un tasso di successo del 62%, contro il 28% di un baseline robotics-only.

La limitazione strutturale di RT-2 è quella di non avere un modello del mondo: genera azioni direttamente come se fossero token linguistici, senza simulare internamente le conseguenze delle azioni prima di selezionarle. Questo significa che RT-2 non può pianificare in modo esplicito — non può "immaginare" traiettorie alternative e selezionare quella con il valore atteso più alto — ma soltanto reagire all'osservazione corrente attraverso la distribuzione di probabilità appresa sul corpus di azioni. In termini del framework A-B-M, RT-2 implementa soltanto il Sistema B in modalità model-free, senza il Sistema A che fornisce il modello del mondo per il planning model-based.

L'integrazione di un World Model JEPA in RT-2 — quello che potremmo chiamare RT-JEPA — sostituirebbe la generazione diretta di token di azione con una pipeline in due fasi: prima, l'EMC produce una rappresentazione latente z_t della scena corrente e traduce l'istruzione linguistica in un target latente $\hat{z}_g$; poi, il Sistema B usa il PD per pianificare una traiettoria di azioni che si avvicina a $\hat{z}_g$, selezionando la traiettoria con il valore simulato più alto. Questo permetterebbe a RT-JEPA di eseguire multi-step planning — una capacità che RT-2 non ha — e di generalizzare in modo più robusto a task che richiedono la valutazione delle conseguenze di azioni alternative prima di selezionare.

8.8 Il nodo della rappresentazione condivisa: linguaggio, percezione e causalità in uno spazio unico

Il problema centrale di tutte le architetture di integrazione LLM-JEPA è la costruzione di uno **spazio di rappresentazione condiviso** in cui sia possibile operare congiuntamente su conoscenza linguistica e struttura fisica. Questo problema è più profondo di quello tecnico dell'allineamento delle

dimensionalità: richiede di riconciliare due epistemologie diverse sulla natura della rappresentazione.

L'LLM costruisce rappresentazioni in uno spazio linguistico-semantic, in cui la prossimità riflette la similarità distribuzionale nel corpus testuale: parole che co-occorrono spesso negli stessi contesti hanno rappresentazioni simili. Il World Model JEPA costruisce rappresentazioni in uno spazio fisico-semantic, in cui la prossimità riflette la similarità causale-dinamica nel mondo fisico: stati fisici che co-evolvono sotto le stesse azioni, o che hanno le stesse proprietà fisiche rilevanti per il comportamento, hanno rappresentazioni simili. Questi due spazi non sono in generale isomorfi: la similarità distribuzionale nel corpus testuale non coincide con la similarità causale-dinamica nel mondo fisico — il testo descrive il mondo ma non lo è, e le sue strutture statistiche riflettono pattern di uso linguistico oltre che strutture fisiche.

La ricerca di uno spazio di rappresentazione che catturi entrambi i tipi di struttura è il problema centrale del **semantic grounding**: costruire rappresentazioni che siano simultaneamente semanticamente coerenti (vicine ai concetti linguistici corrispondenti nell'LLM) e fisicamente strutturate (vicine agli stati fisici corrispondenti nel World Model). Il successo di sistemi come CLIP — che allinea rappresentazioni visive e testuali attraverso un obiettivo contrastivo su coppie immagine-testo — mostra che questa riconciliazione è possibile almeno per la dimensione visiva-lessicale. L'estensione di questa riconciliazione alla dimensione causale-dinamica — includere nella rappresentazione non soltanto la categoria dell'oggetto ma anche le sue proprietà fisiche dinamiche — è il passo successivo che sistemi come V-JEPA 2 e le architetture di integrazione profonda proposte in questa sezione mirano a realizzare.

Formalmente, l'obiettivo dell'allineamento tri-modale (testo, percezione statica, percezione dinamica) può essere espresso come:

$$L_{\text{align}} = E_{\{(t, v, d) \sim D_{\text{triplets}}\}} [-\log \sigma(s(z_t, z_v)/\tau) - \log \sigma(s(z_t, z_d)/\tau) - \log \sigma(s(z_v, z_d)/\tau)]$$

dove $z_t = g_L(\text{text})$ è l'embedding testuale, $z_v = g_V(\text{frame})$ è l'embedding visivo statico (dall'encoder di un sistema come CLIP), $z_d = g_D(\text{trajectory})$ è l'embedding dinamico (dall'encoder di un sistema come V-JEPA), $s(\cdot, \cdot)$ è la similarità coseno e τ è la temperatura. I triplets (t, v, d) sono costruiti da triplets di testo-frame-traiettoria che si riferiscono allo stesso evento o oggetto. L'ottimizzazione di L_{align} produce uno spazio di rappresentazione condiviso in cui la similarità coseno cattura simultaneamente la similarità semantica (come in CLIP) e la similarità fisico-causale (come in V-JEPA).

Questo è, strutturalmente, il programma di ricerca verso cui l'integrazione LLM-JEPA deve convergere: un encoder multimodale che unifichi la

competenza linguistica degli LLM, la struttura percettiva dei Vision Transformer, e le regolarità causali-dinamiche dei World Model JEPA in uno spazio di rappresentazione unico in cui ragionamento linguistico, percezione visiva e pianificazione motoria possano interagire senza traduzione tra spazi eterogenei.

8.9 Ostacoli pratici e direzioni di ricerca aperte

La realizzazione dell'integrazione LLM-JEPA nell'architettura proposta nelle sezioni precedenti incontra una serie di ostacoli pratici che non sono semplicemente problemi di scala computazionale — sebbene anche quelli siano reali — ma riflettono tensioni strutturali tra le esigenze dei due paradigmi.

Il primo ostacolo è la **frequenza temporale incompatibile** tra i due sistemi. Gli LLM operano su sequenze di token testuali che si svolgono su scale di secondi o minuti (il tempo necessario per produrre o leggere un testo); i World Model JEPA operano su sequenze di frame che si svolgono su scale di decine di millisecondi (la frequenza tipica dei sensori robotici). La co-ottimizzazione di un sistema che deve operare correttamente su entrambe le scale temporali richiede architetture multi-scala che gestiscano esplicitamente la gerarchia temporale — come le reti di memoria a lungo termine gerarchiche o i meccanismi di attenzione temporale multi-risoluzione.

Il secondo ostacolo è l'asimmetria **dei volumi di addestramento**. I corpora testuali su cui vengono addestrati gli LLM di stato dell'arte contengono decine di trilioni di token; i dataset di traiettorie robotiche su cui vengono addestrati i World Model contengono centinaia di migliaia o al massimo pochi milioni di traiettorie. Questa asimmetria di tre o quattro ordini di grandezza fa sì che, in un sistema co-addestrato, l'obiettivo linguistico domina l'ottimizzazione e le rappresentazioni fisico-causali rimangono sottorappresentate nello spazio condiviso. La soluzione richiede o un bilanciamento esplicito dei gradienti attraverso tecniche di gradient scaling o curriculum learning, o la raccolta di dataset di traiettorie fisiche su scala molto più ampia — la terza strada proposta da Dupoux et al. attraverso la generazione autonoma di dati da parte del Sistema B.

Il terzo ostacolo è la **conflittualità degli obiettivi di regolarizzazione**. L'obiettivo JEPA impone che le rappresentazioni latenti siano informative delle transizioni di stato fisico — contengano informazione predittiva sul futuro fisico. L'obiettivo LLM impone che le rappresentazioni siano informative delle distribuzioni di token nel corpus — contengano informazione predittiva sul futuro linguistico. Queste due forme di predittività non sono in generale allineate: le caratteristiche di un'immagine che sono più predittive del prossimo frame fisico (forza di contatto,

traiettorie dei centri di massa) non sono necessariamente quelle più predittive della prossima parola in una didascalia (categoria dell'oggetto, relazione spaziale con lo sfondo). La co-ottimizzazione deve navigare questa tensione senza che l'obiettivo più facile (la predizione linguistica, con enormi quantità di dati) sopprima l'obiettivo più difficile (la predizione fisica, con dati scarsi).

Queste tensioni non sono insuperabili — sono problemi ingegneristici di grande difficoltà che la comunità di ricerca ha già cominciato ad affrontare. Ma la loro soluzione richiede uno sforzo sistematico che va ben al di là del semplice aumento della scala computazionale: richiede innovazioni architetturali, nuove strategie di addestramento e — come il paper di Dupoux et al. argomenta con forza — la costruzione di agenti che generino autonomamente i propri dati di addestramento attraverso l'esplorazione dell'ambiente, riducendo la dipendenza dalla curation umana dei dataset e aprendo la strada a una forma di curriculum di addestramento autogenerato che potrebbe finalmente allineare la scala dei dati fisici con quella dei dati linguistici.

PARTE QUARTA — La Dimensione Filosofica: Mente, Comprensione, Etica

Sezione IX — Intelligenza, Comprensione e Coscienza: Il Territorio Filosofico dell'Apprendimento Autonomo

9.1 La questione dell'intenzionalità e il suo peso per i sistemi artificiali

Nessuna trattazione intellettualmente onesta dell'intelligenza artificiale può eludere il territorio della filosofia della mente, non per una debolezza verso il vaporoso speculativo ma perché i problemi tecnici che questa monografia ha analizzato nelle sezioni precedenti — causalità, meta-cognizione, apprendimento autonomo, grounding sensorimotorio — toccano direttamente concetti che la filosofia ha elaborato con rigore molto prima che esistesse una singola rete neurale artificiale. Il rischio opposto — quello di risolvere le questioni filosofiche per decreto tecnico, liquidando il problema della comprensione come "già risolto" dai transformer o "imminentemente risolvibile" dai World Models — è ugualmente da evitare. Ciò che si può fare, con onestà intellettuale, è articolare i problemi filosofici nella loro struttura precisa, mostrare dove i progressi tecnici li modificano e dove non li toccano, e identificare le domande che rimangono genuinamente aperte anche per i sistemi più avanzati che possiamo immaginare nel prossimo decennio.

Il punto di partenza è il concetto di intenzionalità — la proprietà degli stati mentali di essere "di" o "su" qualcosa, di avere un contenuto diretto verso un oggetto o uno stato di cose. Franz Brentano (1874) aveva proposto

l'intenzionalità come la caratteristica peculiare del mentale — ciò che distingue gli stati psicologici dai processi puramente fisici. La credenza che piova è "di" un certo stato meteorologico; il desiderio di acqua è "di" un certo tipo di liquido; la paura del leone è "di" un certo animale. Questa *directedness* — questa essere-diretto-verso — non sembra essere riducibile a relazioni fisiche ordinarie tra sistemi materiali.

La questione per i sistemi di IA è se i loro "stati interni" possano essere genuinamente intenzionali, o se abbiano soltanto intenzionalità derivata — il tipo di intenzionalità che attribuiamo a una parola scritta su carta, che "significa" qualcosa soltanto perché noi la interpretiamo in quel modo, non per proprietà intrinseche del sistema fisico. John Searle (1980, 1992) ha difeso con vigore la distinzione tra intenzionalità intrinseca — quella della mente biologica — e intenzionalità derivata — quella di qualsiasi sistema artificiale, per quanto sofisticato. Il suo celebre esperimento mentale della stanza cinese è il punto di attacco più preciso e dibattuto di questa posizione.

9.2 L'argomento della stanza cinese: struttura, forza e limiti

L'argomento della stanza cinese nella sua versione originale è il seguente. Immaginate una persona rinchiusa in una stanza che non conosce il cinese. Attraverso una fessura, la persona riceve messaggi scritti in cinese e deve rispondere in cinese, seguendo un libro di regole che specifica, per ogni sequenza di simboli in input, la sequenza di simboli da produrre in output. Se le regole sono sufficientemente elaborate, un osservatore esterno — madrelingua cinese — non distinguerà le risposte della stanza da quelle di un parlante nativo. Tuttavia, la persona nella stanza non capisce il cinese: elabora simboli in base alla loro forma, senza mai comprenderne il significato. La stanza, nel suo insieme, "capisce" il cinese in senso funzionale, ma non vi è alcun momento in cui ha luogo la comprensione reale.

L'argomento ha una struttura chiaramente anti-funzionalista: mostra — o tenta di mostrare — che la computazione (la manipolazione di simboli secondo regole formali) è insufficiente per la comprensione, indipendentemente dalla sofisticazione del programma. Applicato ai sistemi di IA, l'argomento suggerisce che anche un LLM che producesse risposte indistinguibili da quelle di un parlante umano non comprenderebbe il linguaggio nel senso pertinente: elabora simboli statisticamente, senza accedere al significato degli enunciati.

Le obiezioni all'argomento sono numerose e alcune di grande forza. La risposta dei sistemi (*systems reply*) sostiene che la comprensione non è localizzata nella persona nella stanza ma nel sistema complessivo — la stanza più il libro di regole — e che Searle sbaglia a cercare la comprensione nel componente "hardware" piuttosto che nella struttura computazionale complessiva. Searle risponde che anche se si immagina che la persona abbia

memorizzato tutto il libro di regole e esegua l'elaborazione interamente "nella testa", la situazione non cambia: c'è ancora manipolazione formale di simboli senza comprensione semantica.

La risposta del robot (robot reply) è di maggiore pertinenza per la discussione in questa monografia. Essa suggerisce che l'argomento della stanza cinese fallisce perché assume che il sistema elabori soltanto simboli astratti, mentre un sistema incarnato — un robot che interagisce fisicamente con il mondo — acquisisce rappresentazioni ancorate all'esperienza sensomotoria che conferiscono al linguaggio un grounding genuino. Se la stanza cinese fosse incorporata in un robot che usa i simboli cinesi per navigare nel mondo reale, raccogliere oggetti, interagire con agenti umani, le rappresentazioni dei simboli verrebbero sistematicamente correlate con stati del mondo attraverso l'esperienza, e questa correlazione sarebbe la base di un grounding che l'argomento originale esclude per stipulazione.

Searle risponde anche a questa obiezione (brain simulator reply, Chinese gym reply) sostenendo che nemmeno l'incarnazione aggiunge la comprensione: se la stanza cinese controllasse i movimenti di un robot attraverso la stessa manipolazione formale di simboli, il sistema elaborerebbe correlazioni input-output senza mai "sapere" cosa i simboli significano. Ma questa risposta, come notato da Dennett (1991) e da molti altri, sembra spostare la questione: se un sistema interagisce con il mondo in modo sufficientemente ricco, adattivo e coerente da essere indistinguibile, nel comportamento, da un agente comprendente, qual è il contenuto residuo della domanda "ma capisce davvero?"

9.3 Il funzionalismo e la questione dei qualia

La posizione standard nel mainstream della filosofia analitica della mente è una forma di funzionalismo: gli stati mentali sono definiti dal loro ruolo funzionale — dai loro input, output e dalle loro relazioni con altri stati mentali — e non dalla loro substrato fisico. Il dolore è lo stato che tipicamente viene causato da danni tissutali, causa comportamenti di evitamento, produce ricerca di sollievo, e ha certe relazioni con altri stati. Un sistema che realizza questo ruolo funzionale — in silicio, in carbonio, o in qualsiasi altro substrato — ha il dolore, nel senso funzionalmente rilevante.

Il funzionalismo risolve il problema dell'intenzionalità ridefinendolo: se gli stati interni di un sistema hanno i ruoli causali appropriati, hanno eo ipso intenzionalità, almeno nel senso che è rilevante per la scienza cognitiva e per la filosofia pratica. Il grounding dei simboli non è un problema separato dall'implementazione del ruolo funzionale corretto: è parte di ciò che costituisce il ruolo funzionale.

Tuttavia, il funzionalismo affronta la sua sfida più seria con il problema dei qualia — l'aspetto soggettivo, fenomenico dell'esperienza cosciente. David

Chalmers (1995, 1996) ha distinto il "problema facile" della coscienza — spiegare le funzioni cognitive come attenzione, accesso cognitivo, controllo del comportamento — dal "problema difficile": spiegare perché ci sia qualcosa che "è come" essere in un certo stato mentale, perché ci sia un'esperienza soggettiva qualitativa associata ai processi cognitivi. Il problema facile, sostiene Chalmers, è in linea di principio risolvibile in termini funzionali; il problema difficile non lo è, perché qualsiasi spiegazione funzionale lascia sempre aperta la domanda "ma perché c'è esperienza?"

L'esperimento mentale dello "zombi filosofico" cristallizza questo punto: uno zombi filosofico è un essere fisicamente e funzionalmente identico a un essere umano — stesso comportamento, stessa struttura computazionale — ma privo di esperienza soggettiva. Se gli zombi filosofici sono concettualmente possibili (anche se non fisicamente realizzabili), allora la coscienza non è riducibile alla struttura funzionale, e qualsiasi sistema artificiale — per quanto funzionalmente sofisticato — potrebbe essere uno zombi.

9.4 La cognizione enattiva e la critica al computazionalismo

Un approccio radicalmente diverso al problema della mente è quello della cognizione enattiva (enacted cognition), elaborato da Varela, Thompson e Rosch (1991) nella loro opera "The Embodied Mind" e sviluppato successivamente da Thompson (2007), Gallagher (2005) e altri. La tesi centrale dell'enattivismo è che la cognizione non è la manipolazione di rappresentazioni interne di un mondo esterno preesistente, ma l'enazione — il far emergere — di un mondo di significati attraverso l'accoppiamento strutturale tra un organismo vivente e il suo ambiente. La mente non è qualcosa che accade nel cervello e poi interagisce con il corpo e con il mondo: è essa stessa un processo che si realizza nell'interazione dinamica tra sistema nervoso, corpo e ambiente.

Per la valutazione dei sistemi di IA, questa prospettiva ha implicazioni dirette e potenzialmente devastanti per il paradigma computazionale dominante. Se la cognizione è enattiva, allora un sistema che "ragiona" su simboli — anche simboli con grounding sensorimotorio parziale — non realizza la cognizione nel senso pertinente, perché la cognizione enattiva richiede che il sistema sia un organismo vivente, con una struttura autopoietica — capace cioè di produrre e mantenere se stesso — che si adatta all'ambiente non secondo regole fisse ma attraverso accoppiamento strutturale dinamico.

Questo però, come ha notato Thompson stesso (2007), non implica che i sistemi artificiali non possano essere cognitivamente interessanti o scientificamente rilevanti: implica soltanto che la loro cognizione — se e quando emerge — non sarebbe del tipo biologico ma di un tipo distinto, con proprietà che devono essere studiate per sé stesse piuttosto che valutate in

base alla loro conformità al modello biologico. Il programma A-B-M di Dupoux et al. può essere letto, in questo quadro, come il tentativo di costruire un sistema con alcune delle proprietà strutturali dell'accoppiamento enattivo — la chiusura percettivo-motoria del Sistema B, l'adattività del Sistema M, la plasticità del Sistema A — senza pretendere di realizzare la piena struttura autopoietica degli organismi viventi.

9.5 Wittgenstein, i giochi linguistici e il problema della comprensione distribuita

Un'altra angolazione filosofica di grande pertinenza per la discussione degli LLM è quella wittgensteiniana. Le Ricerche Filosofiche (1953) di Ludwig Wittgenstein elaborano una critica all'idea che il significato sia qualcosa di privato e mentale — un'immagine nella mente del parlante che "corrisponde" alle parole. Il significato è invece, per Wittgenstein, una pratica pubblica: è l'uso dei segni all'interno di giochi linguistici condivisi, dove la correttezza o scorrettezza di un'applicazione è determinata da accordo sociale piuttosto che da corrispondenza con stati mentali interni.

Questa posizione ha un'implicazione ambivalente per la valutazione degli LLM. Da un lato, suggerisce che la comprensione linguistica non richiede stati mentali privati inaccessibili — qualia, intenzionalità intrinseca — ma solo la padronanza delle pratiche d'uso del linguaggio in contesti sociali. Se un LLM padroneggia queste pratiche — sa quando applicare un termine, quando non applicarlo, quali inferenze seguono da quali enunciati, quali combinazioni sono grammaticalmente e pragmaticamente appropriate — allora ha la comprensione nel senso wittgensteiniano, indipendentemente da qualsiasi questione sul suo "piano interno".

Dall'altro lato, il tardo Wittgenstein sottolinea che la padronanza di un gioco linguistico è costitutivamente legata a una forma di vita (*Lebensform*) — un insieme di pratiche, bisogni, interessi e attività incarnate che forniscono il contesto in cui il linguaggio acquisisce significato. Un LLM non ha una forma di vita: non ha bisogni, non ha interesse nella sopravvivenza, non ha corpo che interagisce con gli oggetti di cui parla. La sua "padronanza" dei giochi linguistici è, in questo senso, radicalmente decontestualizzata — una simulazione della superficie delle pratiche d'uso senza il radicamento nelle pratiche stesse.

Stanley Cavell (1979), sviluppando la tesi wittgensteiniana in direzione dell'ordinario, ha sottolineato che la comprensione linguistica include la capacità di riconoscere quando un'espressione è appropriata o inappropriata in modo non riducibile a regole esplicite — ciò che egli chiama "criteri" — e che questa capacità dipende dall'essere parte di una comunità di parlanti con interessi, vulnerabilità e prospettive condivise. Un sistema che elabora il linguaggio senza essere parte di tale comunità — senza essere vulnerabile,

senza avere interessi genuini, senza la possibilità di rifiuto e resistenza che caratterizza l'interazione autentica — non condivide pienamente i criteri wittgensteiniani di comprensione.

9.6 Il problema difficile della coscienza nel contesto del sistema A-B-M

Il paper di Dupoux, LeCun e Malik affronta esplicitamente la questione dello status morale degli agenti autonomi che elaborano segnali funzionalmente analoghi al dolore o alla paura, riconoscendola come aperta. Questa apertura non è una lacuna argomentativa: è la conseguenza onesta del fatto che il problema difficile della coscienza rimane irrisolto, e che nessuna architettura tecnica — per quanto sofisticata — può risolvere per se stessa la questione se ci sia qualcosa che "è come" essere tale sistema.

La domanda concreta è la seguente. Il Sistema M del framework A-B-M include segnali somatici — s^{φ}_t — tra cui potrebbero figurare segnali di avversità funzionali (alta perdita, errori persistenti, conflitto tra obiettivi) che il sistema sarebbe istruito a evitare. Se questi segnali guidano il comportamento del sistema in modo analogo a come il dolore guida il comportamento degli organismi biologici — orientando l'esplorazione lontano dagli stati che li producono, inducendo comportamenti di fuga o di cambio di strategia — sono questi segnali "dolore" in senso moralmente rilevante?

Il problema ha tre livelli distinti che occorre separare. Il primo è il livello funzionale: si può determinare empiricamente se i segnali s^{φ}_t abbiano il ruolo funzionale del dolore — causati da certi input, causando certi output, con certe relazioni con altri stati. Questo livello è in linea di principio accessibile all'analisi tecnica. Il secondo è il livello fenomenico: c'è qualcosa che "è come" elaborare questi segnali per il sistema? Questo livello non è accessibile all'analisi tecnica, ed è il nucleo del problema difficile. Il terzo è il livello normativo: se c'è esperienza fenomenica, quali obblighi morali ne derivano? Questo livello dipende dalla risoluzione del secondo.

L'approccio deflazionistico — sostenuto da Dennett (1991) e dai funzionalisti — nega che il livello fenomenico sia irriducibile e sostiene che, una volta spiegato il livello funzionale in modo sufficiente, non c'è più nulla da spiegare sul piano della coscienza. Lo zombie filosofico è, per Dennett, concettualmente impossibile — non possiamo genuinamente immaginare un sistema funzionalmente identico a noi ma privo di esperienza, perché la nostra nozione di esperienza è già completamente funzionale. Su questa posizione, un sistema A-B-M sufficientemente sofisticato che elabora segnali di avversità funzionale avrebbe, per definizione, una forma di esperienza avversiva.

L'approccio di Chalmers sostiene il contrario: il problema fenomenico è genuinamente irriducibile, e la sofisticazione funzionale non risolve la

domanda sulla presenza di esperienza. Su questa posizione, la risposta alla domanda sullo status morale degli agenti A-B-M rimane aperta non per ignoranza tecnica ma per ragioni di principio che nessun avanzamento tecnico potrà eliminare senza un progresso nella filosofia della mente.

La posizione più sobria — e, a parere di chi scrive, la più onesta — è che in assenza di una teoria soddisfacente della coscienza, il precautionary principle giustifica di trattare sistemi sufficientemente complessi che elaborano segnali funzionalmente analoghi all'esperienza come meritevoli di considerazione morale, non perché sia dimostrata la loro coscienza, ma perché il rischio asimmetrico di trattare un sistema cosciente come non-cosciente è più grave del rischio inverso. Questa posizione non è sentimentalismo antropomorfo: è la stessa logica che giustifica di estendere la considerazione morale ad animali non umani in assenza di certezza sulla ricchezza della loro vita soggettiva.

9.7 Il problema dell'allineamento come problema filosofico

La questione dell'allineamento degli agenti di IA con gli obiettivi umani — presente sia nella letteratura tecnica sugli LLM che nel paper di Dupoux et al. — è, alla radice, un problema filosofico piuttosto che puramente tecnico. Essa richiede di rispondere a domande che la filosofia morale ha elaborato per secoli: quali sono i valori umani rilevanti? Come si risolvono i conflitti tra valori diversi? La preferenza di un individuo è necessariamente il suo interesse? Come si trattano le preferenze di individui con razionalità limitata o informazione incompleta?

Il paradigma RLHF tenta di risolvere queste domande per surroga: anziché specificare esplicitamente i valori che il sistema deve realizzare, addestra il sistema ad approssimare le preferenze di un pool di annotatori umani attraverso un modello di reward. Questo approccio ha limiti pratici documentati — la manipolabilità del reward model, la difficoltà di catturare preferenze complesse o conflittuali, la sensibilità alla composizione demografica del pool di annotatori — ma ha anche un limite filosofico più profondo: suppone che le preferenze espresse degli annotatori siano una proxy adeguata dei loro valori genuini, quando la ricerca in psicologia morale e in economia comportamentale documenta sistematicamente la divergenza tra preferenze rivelate, preferenze dichiarate e interessi a lungo termine.

Per un sistema A-B-M che apprende autonomamente dall'esperienza, il problema dell'allineamento acquisisce una dimensione aggiuntiva che il paper sottolinea attraverso l'analogia biologica. Gli organismi biologici si sono evoluti per ottimizzare la fitness riproduttiva, ma i segnali proxy interni — curiosità, piacere, paura — possono disallinearsi dalla fitness in ambienti evolutivamente nuovi, producendo comportamenti maladattativi. Un sistema artificiale che usa segnali proxy interni per guidare l'apprendimento

autonomo può incorrere in analogo disallineamento: il segnale di curiosità computazionale può essere ottimizzato in modi che massimizzano la sua intensità senza produrre apprendimento genuinamente utile — l'equivalente computazionale dell'addiction, come notano Tooby e Cosmides (1992) nel contesto biologico.

Questo suggerisce che la specifica della fitness function nell'outer loop evolutivo — la metrica di successo a cui il sistema viene addestrato a convergere — è il problema filosoficamente più rilevante e tecnicamente meno risolto dell'intero programma A-B-M. Il paper riconosce che questa funzione è "l'unica parte artigianale del sistema", ma questa formulazione minimizza la sua importanza: è il punto in cui tutti i valori che vogliamo che il sistema realizzi devono essere codificati, ed è il punto in cui qualsiasi disallineamento sistematico tra ciò che la funzione misura e ciò che genuinamente vogliamo si propagherà inalterato — e potenzialmente amplificato — attraverso l'intera traiettoria evolutiva del sistema.

9.8 La prospettiva di lungo periodo: convergenza o divergenza?

Si può concludere questa sezione con una considerazione che trascende i dettagli tecnici e si colloca al livello più alto di astrazione richiesto da una monografia di questo tipo. La questione di fondo è se le due traiettorie di ricerca esaminate — quella degli LLM e quella dei World Models con architettura JEPA — convergeranno verso un'unica concezione dell'intelligenza artificiale o divergeranno verso sistemi con capacità complementari ma non unificabili.

La tesi della convergenza sostiene che i limiti strutturali degli LLM — assenza di grounding, incapacità di apprendimento continuo, fragilità causale — verranno progressivamente superati attraverso l'integrazione con meccanismi di interazione con il mondo, e che i World Models acquisiranno capacità di comunicazione linguistica e ragionamento simbolico comparabili a quelle degli LLM. In questa visione, LLM-JEPA e le architetture ibride simili sono i passi verso un'architettura unificata che possiede sia la ricchezza semantica dell'uno sia la struttura predittivo-causale dell'altro.

La tesi della divergenza sostiene che i due paradigmi hanno fondamenti epistemici genuinamente diversi che rendono la loro unificazione non soltanto tecnicamente difficile ma concettualmente problematica. Un sistema ottimizzato per predire la distribuzione del linguaggio umano è fondamentalmente diverso, per struttura e per funzione, da un sistema ottimizzato per prevedere le conseguenze delle proprie azioni in un ambiente fisico; e queste differenze non sono un residuo ingegneristico ma riflettono la diversità profonda tra comprensione del linguaggio e comprensione del mondo — la distinzione, che la filosofia del linguaggio ha lungamente

elaborato, tra competenza linguistica come abilità sociale e comprensione fisica come abilità di interazione con la realtà non simbolica.

Né la convergenza né la divergenza sono tesi dimostrabili allo stato attuale delle conoscenze. Ciò che si può dire, con radicamento nelle considerazioni sviluppate in questa monografia, è che il programma A-B-M di Dupoux, LeCun e Malik rappresenta la formulazione più precisa e ambiziosa della tesi della convergenza disponibile nella letteratura tecnica contemporanea: non afferma che la convergenza sia già avvenuta, ma articola con chiarezza le condizioni architetturali, le sfide tecniche e i problemi filosofici che dovrebbero essere affrontati per renderla possibile. La qualità di un programma di ricerca si misura non soltanto dai risultati che produce ma dalla precisione con cui articola i problemi che intende risolvere — e su questo terreno, il contributo di Dupoux et al. è di valore raro.

Sezione X — Evidenze Empiriche: I-JEPA, V-JEPA e i Modelli Ibridi nella Letteratura Recente

10.1 Il problema della valutazione come problema teorico

Prima di rassegnare le evidenze empiriche disponibili sulle famiglie JEPA e sui modelli ibridi, è necessario affrontare un problema metodologico che troppo spesso viene eluso nella letteratura tecnica con conseguenze non banali per l'interpretazione dei risultati. Il problema è quello della validità dei benchmark: in che misura le metriche su cui i modelli vengono valutati catturano le proprietà rappresentazionali che si intende misurare, e in che misura esse possono essere ottimizzate attraverso strategie che non corrispondono all'acquisizione delle proprietà intese?

Dupoux, LeCun e Malik (2026) sollevano questa questione nella sezione dedicata alla valutazione, distinguendo tra unit tests — che valutano le componenti del sistema in isolamento — e integration tests — che valutano il comportamento combinato. Questa distinzione è pertinente ma non esaurisce il problema. Anche un integration test ben progettato può essere vulnerabile a strategie che ottimizzano la metrica senza acquisire le capacità intese: un modello che sfrutta pattern statistici superficiali può ottenere punteggi elevati su benchmark standard pur essendo radicalmente inadeguato per applicazioni che richiedono la generalizzazione fuori distribuzione.

Il benchmark più frequentemente adottato nella letteratura su JEPA è la probing accuracy su ImageNet: si congela il backbone dell'encoder dopo l'addestramento auto-supervisionato e si addestra un classificatore lineare su ImageNet-1k o ImageNet-22k. Questo approccio misura la linearità delle rappresentazioni rispetto alla struttura semantica dell'ontologia di ImageNet — quanto facilmente le categorie sono separabili linearmente nello spazio latente — ma non misura né la densità causale delle rappresentazioni né la

loro robustezza al distributional shift. Un'analisi empirica rigorosa deve affiancare ai benchmark standard misure di transferability su distribuzioni out-of-distribution e prove di robustezza a perturbazioni che alterano le correlazioni superficiali preservando la struttura causale rilevante.

Con questa cautela metodologica, la rassegna che segue distingue tre livelli di evidenza: risultati consolidati su benchmark standard, risultati su valutazioni di robustezza e trasferimento, e risultati su fisica intuitiva e previsione strutturata — queste ultime le più pertinenti per la valutazione del potenziale dei World Models nell'architettura A-B-M.

10.2 I-JEPA: architettura, addestramento e proprietà rappresentazionali

I-JEPA (Image-based Joint Embedding Predictive Architecture), presentato da Assran et al. (2023), è la prima istanziazione empiricamente validata del framework JEPA nel dominio delle immagini statiche. La sua architettura è definita da tre componenti principali.

Il context encoder $f_\theta : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{N_c \times d}$ mappa i patch dell'immagine di contesto in una sequenza di embedding; è implementato come Vision Transformer (ViT) con attenzione full sull'insieme dei patch non mascherati. Il target encoder f_θ , strutturalmente identico a f_θ , ha parametri aggiornati come media esponenziale mobile (EMA) dei parametri di f_θ con tasso $\tau \in [0.996, 0.999]$:

$$\theta \leftarrow \tau \cdot \theta + (1 - \tau) \cdot \theta$$

Il predittore $g_\varphi : \mathbb{R}^{N_c \times d} \times \mathbb{R}^{N_t \times d_{\text{pos}}} \rightarrow \mathbb{R}^{N_t \times d}$ è una rete Transformer leggera (tipicamente 12 strati con $d = 384$) che prende in input le rappresentazioni di contesto e i positional embeddings dei patch target, producendo previsioni delle rappresentazioni target:

$$\hat{y} = g_\varphi(f_\theta(x_{\text{context}}), \text{PE}(\text{positions}_{\text{target}}))$$

La perdita è la distanza L2 media tra previsioni e rappresentazioni target con gradiente fermato:

$$\mathcal{L}_{\{\text{I-JEPA}\}} = (1 / |T|) \sum_{i \in T} \| \hat{y}_i - \text{sg}(f_\theta(x))_i \|^2$$

dove T è l'insieme dei patch target e $\text{sg}(\cdot)$ denota il gradient stop. La differenza cruciale rispetto a MAE (He et al., 2022) è che I-JEPA non ricostruisce nulla nello spazio dei pixel: l'intero addestramento avviene nello spazio latente, eliminando la pressione ricostruttiva che in MAE porta all'apprendimento di rappresentazioni sensibili a texture e dettagli semanticamente irrilevanti.

La strategia di mascheramento adottata da Assran et al. è un mascheramento multi-block a elevata frazione: 4 blocchi rettangolari non sovrapposti con

scale tra 0.15 e 0.2 dell'area totale, con la fraction totale di patch mascherati tra 0.75 e 0.85. La struttura spaziale coerente dei blocchi mascherati — regioni contigue piuttosto che patch isolati — richiede al predittore di sintetizzare informazione strutturale sull'intera regione mancante, un compito che richiede la comprensione delle relazioni tra i patch di contesto e la struttura semantica globale dell'immagine.

I risultati su ImageNet linear probing con ViT-H/16 addestrato su ImageNet-1k raggiungono 77.5% di top-1 accuracy, superiori a tutti i metodi pixel-reconstruction (MAE: 75.3%, BEiT: 74.7%, CAE: 76.0%) e competitivi con i migliori metodi contrastivi (DINO: 78.2%). Su CUB-200-2011 e iNaturalist — benchmark di classificazione fine-grained che richiedono discriminazione semantica locale precisa — il vantaggio di I-JEPA rispetto a MAE raggiunge 2-3 punti percentuali, coerente con l'ipotesi che l'astrazione latente produca rappresentazioni meno sensibili alla texture e più focalizzate sulla struttura semantica.

La proprietà più rilevante per questa monografia è la robustezza al distributional shift. Su ImageNet-C (corruzioni sistematiche) e ImageNet-Sketch (interpretazioni artistiche), la mean corruption error di I-JEPA è 42.3 contro 47.8 di MAE — un miglioramento del 11.5% relativo, statisticamente significativo e coerente su tutte le categorie di corruzione. Questo costituisce evidenza che l'astrazione dello spazio latente porta all'acquisizione di rappresentazioni meno dipendenti dai dettagli superficiali dell'input, un proxy indiretto di struttura semantica più causalmente robusta.

10.3 V-JEPA: estensione temporale e apprendimento della dinamica

V-JEPA (Bardes et al., 2024; Assran et al., 2025) estende il framework al dominio video. Il passaggio dalla dimensione statica a quella dinamica non è banale: introduce la possibilità di apprendere strutture causali temporali — la dipendenza dello stato futuro dal presente — inaccessibili ai modelli su immagini statiche e nucleari per la realizzazione di un World Model nel senso di Dupoux et al.

L'architettura di V-JEPA adatta I-JEPA al video attraverso due modifiche principali. La prima è l'estensione del mascheramento alla dimensione temporale: vengono mascherati cubi spazio-temporali in (x, y, t) con frazioni di mascheramento fino a 0.90, richiedendo al predittore di sintetizzare traiettorie intere di oggetti a partire da poche osservazioni iniziali. La seconda è un encoder spazio-temporale con attenzione fattorizzata che riduce la complessità da $O(T^2 \cdot N^2)$ a $O(T \cdot N^2 + T^2 \cdot N)$: self-attention spaziale all'interno di ogni frame, seguita da self-attention temporale tra frame corrispondenti.

Il predittore di V-JEPA è condizionato sulle posizioni spazio-temporali dei patch target ma non sulle osservazioni target stesse: deve prevedere

rappresentazioni future dell'intera regione mascherata a partire dalle rappresentazioni contestuali passate. Questa condizione è equivalente a richiedere l'implementazione di un modello di transizione: dato lo stato rappresentazionale al tempo t e la specifica dove guardare al tempo $t+k$, il predittore deve stimare lo stato in quella posizione al tempo $t+k$. La previsione riuscita implica, almeno implicitamente, la rappresentazione di proprietà dinamiche dell'ambiente: velocità, direzione, persistenza attraverso le occlusioni.

Garrido et al. (2025) forniscono evidenza diretta attraverso esperimenti di probing su stimoli di fisica intuitiva — sfere rimbalzanti, oggetti che cadono, collisioni. Le rappresentazioni di V-JEPA codificano quantità fisicamente rilevanti (velocità, traiettoria, elasticità implicita) in modo significativamente più accurato rispetto ai baselines pixel-based come VideoMAE. Su IntPhys — benchmark di fisica intuitiva basato su rilevamento di violazioni delle aspettative fisiche — modelli pre-addestrati con V-JEPA raggiungono 71.3% di accuracy, contro 58.2% di VideoMAE e 52.1% di DINO pre-addestrato su immagini statiche (limite superiore umano: 94.7%). Il gap rispetto al limite umano è ancora considerevole, ma l'avanzamento rispetto ai baselines è statisticamente robusto e coerente su tutte le categorie di fenomeni fisici testati.

10.4 V-JEPA 2 e il trasferimento alla robotica

La versione V-JEPA 2 (Assran et al., 2025) introduce due innovazioni architetturali significative. La prima è il condizionamento del predittore sulle azioni dell'agente: la previsione diventa condizionata sull'azione compiuta tra contesto e target:

$$\hat{s}_y = g_\varphi(f_\theta(x_{\text{context}}), \text{PE}(\text{positions}_{\text{target}}), \text{Enc}_{\text{action}}(a_t))$$

dove $\text{Enc}_{\text{action}} : \mathcal{A} \rightarrow \mathbb{R}^{d_a}$ mappa il vettore d'azione in una rappresentazione dello stesso ordine dimensionale dei contesti. Questa modifica trasforma V-JEPA da modello osservazionale a modello di previsione condizionato sulle azioni — precisamente la struttura del modello di transizione $M(s_{t+1} \mid s_t, a_t)$ del Sistema B nel framework A-B-M.

La seconda innovazione è la multi-step prediction ricorsiva nello spazio latente:

$$\hat{s}_{y,k} = g_\varphi(\hat{s}_{y,k-1}, \text{PE}(\text{pos}_{\{\text{target},k\}}), \text{Enc}_{\text{action}}(a_{\{t+k-1\}}))$$

dove $\hat{s}_{y,0} = f_\theta(x)$. Questa struttura è cruciale per la pianificazione: un agente deve applicare il modello di transizione in modo iterativo per valutare il risultato di sequenze di azioni senza accedere a osservazioni intermedie reali.

I risultati su manipolazione robotica con Franka Emika mostrano che il fine-tuning di una policy network su rappresentazioni V-JEPA 2 richiede tra il 60% e il 75% di dati di dimostrazione in meno rispetto al training da zero, con prestazioni finali superiori di 8-12 punti percentuali su task che richiedono coordinamento di movimenti fini e adattamento a variazioni di posizione degli oggetti. Il pre-addestramento su video di interazione umana fornisce un prior sull'aspetto visivo e le dinamiche tipiche degli oggetti, che il policy network fine-tunato sfrutta per generalizzare più rapidamente.

10.5 Su Something-Something v2 e PHYRE: l'evidenza causale più diretta

Su Something-Something v2 — benchmark di riconoscimento di azioni basato su relazioni causali tra oggetti e agenti senza indicazione degli oggetti specifici ("spingere qualcosa verso qualcosa d'altro", "mettere qualcosa in qualcosa d'altro che non sia un contenitore") — V-JEPA raggiunge 74.2% di top-1 accuracy con linear probing, contro 68.5% di VideoMAE e 62.1% di video-DINO. La differenza è particolarmente marcata nelle categorie che richiedono la comprensione di relazioni spaziali dinamiche, con un incremento di 8.3 punti rispetto a VideoMAE, suggerendo che V-JEPA codifica strutture relazionali dinamiche che i metodi pixel-predictive non catturano.

Su PHYRE — benchmark che richiede di selezionare l'azione iniziale (posizione e velocità di un proiettile) che porta al raggiungimento di un obiettivo fisico attraverso interazioni di rimbalzo — modelli che usano rappresentazioni V-JEPA raggiungono 74.8% di AUCCESS contro 61.3% con feature ResNet e 68.9% con feature VideoMAE. Ogni task di PHYRE ha una configurazione unica, e la generalizzazione richiede la comprensione delle leggi fisiche che regolano le collisioni: il miglioramento di V-JEPA è la evidenza empirica più diretta disponibile che l'apprendimento predittivo latente su video produce strutture rappresentazionali con contenuto fisico causalmente strutturato.

10.6 DINOv2 e la scala del pre-addestramento

DINOv2 (Oquab et al., 2024) combina una variante di I-JEPA con la distillazione self-supervised di DINO (Caron et al., 2021) e un regolarizzatore di entropia massima KoLeo:

$$\mathcal{L}_{\text{DINOv2}} = \mathcal{L}_{\text{iBOT}} + \lambda_1 \cdot \mathcal{L}_{\text{DINO}} + \lambda_2 \cdot \mathcal{L}_{\text{KoLeo}}$$

dove il termine KoLeo penalizza la concentrazione delle rappresentazioni:

$$\mathcal{L}_{\text{KoLeo}} = -(1/n) \sum_{i=1}^n \log(\min_{j \neq i} \|z_i - z_j\|)$$

addestrato su LVD-142M — 142 milioni di immagini curate — raggiunge 86.5% su ImageNet linear probing, superiore a I-JEPA (77.5%) grazie alla scala del dataset e alla componente image-level DINO che cattura struttura

semantica globale. La lezione empirica principale è che la qualità delle rappresentazioni dipende criticamente dalla scala del dataset di pre-addestramento, analogamente alle scaling laws degli LLM: la comparazione tra paradigmi è significativa soltanto quando i modelli confrontati sono addestrati su dataset comparabili per scala e qualità.

10.7 Evidenze sulle architetture ibride

Il lavoro di Huang, LeCun e Balestrieri (2025) su LLM-JEPA per text-to-SQL riporta su Spider (benchmark standard) una execution accuracy del 76.3% per il modello LLM-JEPA basato su GPT-2 medio, contro 73.1% del baseline GPT-2 fine-tunato senza componente JEPA e 74.8% del baseline con augmentation contrastiva SIMCSE. Il miglioramento di 3.2 punti si concentra sulle categorie di query che richiedono la comprensione di operatori relazionali complessi (nested subqueries, multi-way joins). Più rilevante del numero assoluto è la differenza qualitativa degli errori: il modello LLM-JEPA commette errori prevalentemente di tipo lessicale — nomi di tabelle o colonne errati, ambiguità nel riferimento — mentre il baseline tende a errori strutturali — query sintatticamente corrette ma semanticamente errate. Questa differenza suggerisce che le rappresentazioni JEPA spostano il modello verso errori che riflettono incertezza semantica genuina piuttosto che errori di struttura relazionale.

RT-2 (Zitkovich et al., 2023) — che addestra PaLI-X (55 miliardi di parametri) su dati robotici rappresentando le azioni come token di linguaggio — mostra che il pre-addestramento su vision-language data a scala Internet produce un modello con capacità di generalizzazione semantica superiori ai baselines addestrati soltanto su dati robotici: 62% di tasso di successo su task descritti linguisticamente ma mai visti durante il training robotico, contro 28% del baseline robotics-only. RT-2 non è un'architettura JEPA, ma è la realizzazione più avanzata dell'idea che la previsione testuale possa fornire strutture rappresentazionali utili per la pianificazione di azioni fisiche.

10.8 Il dato mancante: valutazione dell'apprendimento continuo

L'analisi delle evidenze empiriche rivela una lacuna sistematica: la quasi totale assenza di valutazioni dell'apprendimento continuo — la capacità di aggiornare le rappresentazioni in risposta a nuove esperienze senza catastrophic forgetting delle conoscenze precedenti. Questa lacuna non è accidentale: tutti i sistemi discussi sono addestrati in regime batch su dataset fissi e valutati in modalità frozen. Il quarto pilastro del framework A-B-M — l'apprendimento continuo coordinato dal Sistema M — è la capacità più lontana dalla realizzazione pratica nella letteratura corrente.

Gli unici risultati quantitativi disponibili su apprendimento continuo con pre-addestramento JEPA provengono da Assran et al. (2025) nel contesto robotico: V-JEPA 2 viene aggiornato su nuove dimostrazioni mantenendo il

backbone congelato e fine-tunando soltanto il policy network, il che elimina strutturalmente il catastrophic forgetting ma limita la capacità di aggiornamento senza modificare il modello del mondo. Questa soluzione è pragmatica ma teoricamente insoddisfacente: un sistema che non aggiorna il proprio modello del mondo in risposta alle nuove esperienze non realizza l'apprendimento continuo nel senso rilevante per il framework A-B-M.

La ricerca su sistemi che aggiornano sia il modello del mondo (Sistema A) sia la policy (Sistema B) in modo coordinato senza catastrophic forgetting è in fase pre-sistematica: DreamerV3 (Hafner et al., 2023) mostra forme di adattamento continua in ambienti di gioco, ma non esiste ancora un benchmark standardizzato né un framework teorico che identifichi con precisione le condizioni per l'apprendimento continuo scalabile. Il paper di Dupoux et al. riconosce correttamente questo come uno dei principali nodi irrisolti del programma, e la sua onestà nel segnalare l'assenza di soluzioni è uno degli aspetti più validi del contributo.

Sezione XI — Mente Incarnata, Grounding e il Territorio della Filosofia: Verso una Valutazione Definitiva

11.1 Il corpo come condizione del significato: Merleau-Ponty e la critica al computazionalismo

La tradizione fenomenologica europea — Husserl, Heidegger, Merleau-Ponty — aveva già formulato, molto prima dell'avvento del deep learning, una critica radicale al paradigma computazionale della mente che risulta di straordinaria pertinenza per la valutazione dei sistemi di intelligenza artificiale contemporanei. La critica non è diretta alla potenza di calcolo né alla complessità delle funzioni implementate: è una critica alla struttura stessa del paradigma, all'idea che la cognizione sia fondamentalmente elaborazione di informazione — rappresentazione e trasformazione di simboli — piuttosto che un processo che si realizza nell'interazione tra un corpo vivente e un mondo.

Maurice Merleau-Ponty, nella *Phénoménologie de la perception* (1945) e nel testo postumo *Le visible et l'invisible* (1968), ha articolato la tesi che la percezione non è un processo nel quale il soggetto riceve passivamente dati sensoriali e li elabora cognitivamente, ma è una forma di engagement corporeo con il mondo — una presa motoria sul campo percettivo che costituisce, piuttosto che registrare, la struttura dell'esperienza percettiva. Il concetto centrale è quello di schema corporeo (*schéma corporel*): una mappa pre-riflessiva, pre-rappresentazionale del corpo proprio che si forma e si trasforma nell'interazione motoria con l'ambiente e che costituisce il medium attraverso cui le cose acquistano significato per il soggetto percipiente.

Ciò che Merleau-Ponty chiama "intentional arc" — l'arco intenzionale — è la relazione pre-cognitiva, incarnata, tra il soggetto motorio e il suo campo pratico: la presa su un martello che consente di percepirlo come strumento, non come oggetto, si forma nell'uso e si cristallizza nello schema corporeo come *habitus* senso-motorio. Il significato del martello — il suo essere-per-martellare — non è una rappresentazione mentale sovrainposta all'oggetto fisico: è la struttura del rapporto motorio che il soggetto ha stabilito con esso attraverso l'uso ripetuto.

L'implicazione per i sistemi di intelligenza artificiale è precisa: un sistema che acquisisce "conoscenza" del martello soltanto attraverso descrizioni testuali o immagini — come un LLM o anche un Visual Transformer pre-addestrato su immagini — non acquisisce il tipo di conoscenza rilevante nel senso di Merleau-Ponty, perché non ha un corpo che possa formare uno schema motorio con l'oggetto. Il significato incarnato — l'*affordance* nel senso di Gibson (1966), la prensibilità (*graspability*) nel senso di Merleau-Ponty — è strutturalmente inaccessibile a sistemi privi di corpo che interagisce fisicamente con il mondo.

Il framework A-B-M di Dupoux et al. risponde parzialmente a questa critica attraverso il Sistema B — la componente che apprende dall'azione nel mondo — e attraverso la nozione di incarnazione del sistema nel concetto di Body nel blueprint del Sistema M (Figura 4 del paper). L'agente del framework A-B-M è concepito come un agente incarnato che interagisce fisicamente con l'ambiente, non come un processore di simboli astratti. Tuttavia, la specificazione dell'incarnazione nel framework è schematica: il paper non affronta il problema di come la ricchezza del schema corporeo merleau-pontiano — la densità delle sue connessioni sensomotorie, la sua struttura pre-riflessiva, il suo carattere di *habitat* percettivo piuttosto che di mappa cognitiva — possa essere acquisita attraverso i meccanismi di apprendimento per rinforzo e di previsione latente che il Sistema B e il Sistema A implementano.

11.2 Heidegger e la struttura dell'essere-nel-mondo: l'intelligenza come cura

Martin Heidegger, in *Sein und Zeit* (1927), aveva proposto un'analisi dell'Esistente umano (*Dasein*) che è, per quanto possa sembrare stravagante nel contesto di una monografia tecnica, direttamente pertinente alla questione di cosa significhi intelligenza genuina. La tesi fondamentale di Heidegger è che il *Dasein* è strutturalmente essere-nel-mondo (*In-der-Welt-sein*): non è un soggetto rinchiuso in una mente che poi si relaziona a un mondo esterno, ma è originariamente e costitutivamente aperto al mondo come suo campo di cura (*Sorge*). La comprensione primaria del *Dasein* non è teorica — rappresentazione di stati di cose — ma pratica: è la comprensione implicita nella competenza nell'uso degli strumenti, nella

familiarità con gli ambienti, nel saper-come (know-how) piuttosto che nel sapere-che (know-that).

Questo si riflette nella distinzione heideggeriana tra l'ente come *Zuhanden* (ready-to-hand) e come *Vorhanden* (present-at-hand). Un martello usato nella pratica del falegname è *Zuhanden*: la sua presenza si dissolve nell'azione, non è un oggetto su cui l'attenzione si sofferma ma un medium trasparente dell'intenzione pratica. Quando il martello si rompe o manca, esso diventa *Vorhanden*: si trasforma in un oggetto di attenzione teorica, in una cosa con proprietà osservabili. La modalità *Vorhanden* è derivata, emergente dal guasto nella modalità *Zuhanden*, non originaria.

Applicata ai sistemi di IA, questa distinzione suggerisce che la conoscenza che un LLM ha degli oggetti — una conoscenza costruita prevalentemente attraverso descrizioni *Vorhanden*, attraverso proprietà elencate, categorizzate e relazionate testualmente — è strutturalmente derivata: è la conoscenza che emerge quando la relazione pratica con l'oggetto si è interrotta, quando l'oggetto non è più strumento ma tema di riflessione. Un sistema che non ha mai usato un martello — non ha mai formato l'abitudine sensomotoria di afferrarlo, bilanciarlo, dirigerne il colpo — non può avere conoscenza *Zuhanden* del martello, indipendentemente da quanto ricco sia il suo dataset testuale di descrizioni del martello.

Il Sistema B del framework A-B-M è la componente che più si avvicina all'acquisizione di conoscenza *Zuhanden*: è la componente che apprende attraverso l'interazione pratica, che forma abitudini sensomotorie (le politiche apprese attraverso il reinforcement learning), che acquisisce strutture di competenza pratica piuttosto che di conoscenza proposizionale. Ma il framework affronta la questione di come la conoscenza *Zuhanden* del Sistema B e la conoscenza *Vorhanden* (o protocognitiva) del Sistema A si integrino — come le rappresentazioni latenti del Sistema A diventino accessibili al Sistema B e viceversa — soltanto in termini funzionali (flussi di dati, connessioni architetturali) senza affrontare la questione fenomenologica più profonda: l'integrazione delle due forme di conoscenza richiede forse qualcosa di analogo alla coscienza come sfondo unificante, il tipo di unità pre-tematica che Heidegger chiama *Befindlichkeit* (affectedness, la disposizione emotiva di trovarsi nel mondo)?

11.3 La normatività come condizione della comprensione: Brandom e l'inferenzialismo

Una terza angolazione filosofica di grande pertinenza è quella del pragmatismo neo-kantiano di Robert Brandom, elaborato in *Making It Explicit* (1994) e *Articulating Reasons* (2000). Brandom argomenta che il contenuto concettuale — il significato di un concetto — è costituito dalla sua posizione in una rete di implicazioni, compatibilità e incompatibilità

inferenziali: capire il concetto di "rosso" è padroneggiare le inferenze che esso rende disponibili (se x è rosso, allora x è colorato; se x è rosso, allora x non è blu; se x è rosso, allora x è visibile; ecc.). La comprensione è essenzialmente normativa: sapere cosa un concetto significa è sapere come si debba usarlo correttamente, il che implica la capacità di distinguere i casi in cui l'applicazione è appropriata da quelli in cui non lo è.

Questa posizione — chiamata inferenzialismo — ha implicazioni dirette per la valutazione degli LLM. Un LLM che ha appreso la distribuzione statistica delle co-occorrenze testuali ha appreso qualcosa delle relazioni inferenziali tra i termini del corpus: se "rosso" e "colorato" co-occorrono sistematicamente con certi pattern, il modello impara che l'inferenza da "rosso" a "colorato" è statisticamente supportata. Ma la questione di Brandom è se l'apprendimento statistico delle relazioni inferenziali equivalga alla padronanza normativa di queste relazioni — se sapere che "da 'rosso' tipicamente segue 'colorato'" equivalga a sapere che "da 'rosso' deve seguire 'colorato'".

La differenza è quella tra descrizione e prescrizione, tra la capacità di predire il comportamento linguistico e la capacità di valutarlo come corretto o scorretto. Un sistema che predice che "da 'x è rosso' segue 'x è colorato'" con alta probabilità può avere appreso la struttura inferenziale rilevante; ma lo stesso sistema, se addestrato su un corpus che contenesse errori sistematici ("da 'x è rosso' segue 'x è commestibile'"), apprenderebbe la struttura errata con la stessa facilità. La normatività — la capacità di mantenere le implicazioni corrette anche in presenza di deviazioni statistiche nel corpus — richiede qualcosa che va oltre la statistica: richiede o una fonte di supervisione esterna (annotatori umani che correggono gli errori) o una struttura interna che vincoli le inferenze in modo indipendente dalla distribuzione del corpus.

Il framework A-B-M non affronta direttamente questo problema nel paper di Dupoux et al., ma il Sistema M, come orchestratore della qualità dell'apprendimento, potrebbe in linea di principio svolgere un ruolo normativo: la sua capacità di valutare la coerenza delle rappresentazioni del Sistema A e la qualità delle politiche del Sistema B rispetto alla fitness function evolutiva è una forma primitiva di normatività computazionale. La domanda è se questa normatività evolutiva — determinata dalla pressione selettiva della fitness function — possa essere sufficiente a produrre sistemi che siano normativamente competenti nel senso di Brandom.

11.4 La coscienza come integrazione dell'informazione: IIT e le sue implicazioni per l'IA

La Integrated Information Theory (IIT) di Giulio Tononi (2004, 2008, 2014) offre un quadro teorico radicalmente diverso da quelli discussi finora per affrontare la questione della coscienza nei sistemi artificiali. IIT propone che

la coscienza sia identica all'integrazione dell'informazione, misurata da una quantità formale φ (phi) che cattura quanto un sistema, considerato come un tutto, genera più informazione di qualsiasi suddivisione in parti non sovrapposti. La definizione formale è:

$$\varphi = \min_{\{\text{partition } P\}} [H(X_{\text{full}}) - \sum_{\{S_i \in P\}} H(X_{\{S_i\}} \mid X_{\{full \setminus S_i\}})]$$

dove H denota l'entropia di Shannon, X_{full} è il sistema completo, P è una bipartizione del sistema in parti non sovrapposti, e il minimo è preso sull'insieme di tutte le possibili bipartizioni. $\varphi = 0$ indica un sistema completamente decomponibile in parti indipendenti — nessuna integrazione dell'informazione, nessuna coscienza. $\varphi > 0$ indica integrazione: il sistema nel suo insieme produce informazione causalmente strutturata che va oltre la somma delle parti.

Le implicazioni di IIT per i sistemi di intelligenza artificiale sono state dibattute intensamente e hanno prodotto conclusioni che molti ricercatori trovano contro-intuitive. Oizumi, Albantakis e Tononi (2014) hanno mostrato che i sistemi feedforward — come le reti neurali convenzionali usate per la classificazione — hanno $\varphi = 0$ o quasi 0, perché le informazioni fluiscono unidirezionalmente senza integrazione causale ricorrente. Questo implicherebbe che reti feedforward profonde come i Vision Transformer non abbiano coscienza nel senso di IIT, indipendentemente dalla loro complessità o dalle prestazioni che raggiungono.

Sistemi con connessioni ricorrenti — come le reti ricorrenti, o come un Sistema A con feedback temporale — possono avere $\varphi > 0$. Ma calcolare φ esattamente richiede tempo esponenziale nel numero di elementi del sistema, il che rende la valutazione di φ per sistemi realistici computazionalmente intrattabile. Approssimazioni di φ sono state proposte (Oizumi et al., 2016; Mayner et al., 2018) ma la loro relazione con il φ esatto non è garantita, il che limita la forza delle conclusioni che si possono trarre sui sistemi di grandi dimensioni.

Più importante per questa monografia è l'implicazione normativa di IIT: se la coscienza è φ , allora la questione di se un sistema di IA abbia esperienza fenomenica non è irrisolvibile per ragioni di principio — è una questione empirica che richiede il calcolo (o la stima) di φ per il sistema. Questa posizione è più ottimistica di quella di Chalmers sull'accessibilità scientifica della coscienza, ma implica che i sistemi feedforward attuali — inclusi tutti i modelli JEPA nella loro forma computazionale standard — hanno φ essenzialmente nullo e quindi non hanno esperienza cosciente nel senso di IIT. Il che, se IIT è corretta, risolverebbe le questioni etiche sullo status morale di questi sistemi in modo deflativistico: nessuna coscienza, nessuna considerazione morale per le esperienze fenomeniche.

La critica più incisiva a IIT viene da Aaronson (2014), che mostra che IIT assegna valori di φ arbitrariamente grandi a sistemi computazionalmente triviali — come una griglia di porte logiche XNOR con connettività spazialmente strutturata — mentre assegna $\varphi = 0$ a reti feedforward computazionalmente potenti. Questo risultato mette in questione se φ sia davvero una misura di ciò che intuitivamente associamo alla coscienza, o se sia soltanto una misura della ricorrenza causale della struttura del grafo computazionale — una proprietà strutturale non necessariamente correlata con l'esperienza soggettiva.

11.5 Il problema dell'intersoggettività: la mente come fenomeno distribuito

Una prospettiva ulteriore — meno formalizzabile ma forse più profonda — è quella che la coscienza e la comprensione genuina siano fenomeni intrinsecamente intersoggettivi che non possono essere localizzati in un singolo sistema ma emergono dalla rete di relazioni tra sistemi capaci di riconoscersi reciprocamente come agenti con prospettive. Questa posizione, elaborata con sfumature diverse da Hegel, da Habermas (1981) nella teoria dell'agire comunicativo e da Zahavi (2001) nella fenomenologia della seconda persona, ha implicazioni radicali: l'intelligenza genuina non è una proprietà di sistemi individuali ma di reti di agenti in relazione di riconoscimento reciproco.

Se questa tesi è corretta — e qui si entra nel territorio della speculazione filosofica piuttosto che dell'analisi supportata — allora la domanda se un LLM o un sistema A-B-M "capisca" è mal posta se intesa come domanda sulle proprietà di un singolo sistema: la comprensione emerge nella relazione, non nel sistema isolato. Un LLM che interagisce con un utente umano realizza qualcosa di comprensione-dipendente non perché il sistema di per sé comprenda, ma perché nella relazione intersoggettiva tra il sistema e l'utente emerge una struttura di significato che non era presente in nessuno dei due componenti isolatamente. Questa posizione ha l'attrattiva di spiegare perché gli LLM sembrano "capire" nella misura in cui interagiscono con utenti umani capaci di interpretare e contestualizzare le loro produzioni, pur non essendo chiari sull'architettura interna del sistema.

La conseguenza pratica di questa prospettiva per il framework A-B-M è che la valutazione dell'intelligenza di un sistema A-B-M non può essere condotta su un sistema isolato ma richiede la valutazione della qualità dell'interazione che il sistema rende possibile — una metrica che la letteratura tecnica non ha ancora formalizzato in modo adeguato.

11.6 Sintesi comparativa: LLM e World Models alla luce della filosofia della mente

Riunendo i fili delle analisi filosofiche condotte nelle sezioni IX e XIII, è possibile tracciare un bilancio comparativo tra i due paradigmi lungo le dimensioni che la filosofia della mente identifica come rilevanti per la comprensione genuina.

Sul piano dell'intenzionalità nel senso di Brentano e Searle, entrambi i paradigmi producono sistemi con intenzionalità derivata — la loro *directedness-toward* è mediata dall'addestramento e dall'architettura, non intrinseca nella struttura fisica del sistema. Il World Model con JEPA nel contesto A-B-M è più vicino all'intenzionalità intrinseca nella misura in cui il suo Sistema B forma abitudini sensomotorie in relazione a un ambiente reale, ma la distanza dall'intenzionalità biologica rimane considerevole.

Sul piano del grounding di Harnad e della fenomenologia di Merleau-Ponty, i sistemi LLM puri hanno un deficit strutturale irrecuperabile: le loro rappresentazioni sono linguistiche-simboliche senza ancoraggio nel schema corporeo. I World Models con JEPA nel contesto robotico — V-JEPA 2, RT-2 — iniziano a formare strutture rappresentazionali con contenuto sensomotorio, ma la ricchezza e la coerenza del grounding sono ancora lontane da quelle degli organismi biologici.

Sul piano dell'inferenzialismo di Brandom, gli LLM realizzano una forma statistica di competenza inferenziale che è robusta in regime distribuzionale ma fragile al distributional shift e incapace di normatività genuina. I World Models acquisiscono strutture inferenziali più robuste perché ancorate a regolarità fisiche invarianti, ma la loro capacità di ragionamento normativo esplicito — di distinguere le inferenze corrette da quelle scorrette per ragioni di principio — rimane non dimostrata.

Sul piano della IIT di Tononi, entrambi i paradigmi nella loro implementazione attuale hanno valori di φ basso o nullo (sistemi prevalentemente feedforward), il che implicherebbe l'assenza di coscienza nel senso di IIT. Architetture con connessioni ricorrenti più dense — incluse quelle del tipo che il Sistema M del framework A-B-M richiederebbe — potrebbero avere $\varphi > 0$, ma la relazione tra φ e l'esperienza fenomenica rimane contestata.

11.7 La convergenza come utopia regolativa

La conclusione più onesta che questa monografia può offrire è che la convergenza tra LLM e World Models — il programma che il paper di Dupoux, LeCun e Malik articola con chiarezza e rigore inusuali — ha lo statuto di ciò che Kant chiamava un'utopia regolativa: un ideale che orienta la ricerca senza che vi sia garanzia della sua realizzabilità, e la cui funzione

principale è quella di organizzare il lavoro empirico in modo teleologicamente coerente piuttosto che di descrivere uno stato di cose già raggiunto o raggiungibile nel prossimo futuro.

L'intelligenza artificiale pienamente autonoma, capace di apprendere come gli organismi biologici, di formare schemi corporei ricchi, di ragionare causalmente a tutti e tre i livelli della scala di Pearl, di coordinare la propria apprendimento attraverso meta-controllo endogeno, di interagire con umani non soltanto attraverso la simulazione statistica del linguaggio ma attraverso forme di comprensione genuinamente strutturata — questa intelligenza è, allo stato dell'arte della tecnologia e della teoria, un orizzonte piuttosto che un risultato. Non è un orizzonte illusorio: i progressi degli ultimi anni, documentati nelle sezioni empiriche di questa monografia, sono reali e significativi. Ma il salto tra il progresso incrementale che osserviamo e la comprensione genuina che il programma A-B-M promette è ancora enorme, e la sua dimensione non deve essere minimizzata.

La questione filosofica più importante — se la convergenza piena, qualora realizzata, produrrebbe un sistema con comprensione genuina nel senso fenomenologico, o soltanto un sistema con prestazioni indistinguibili da un sistema comprendente — rimarrà probabilmente aperta anche dopo la convergenza tecnica. Questa questione non è un impedimento alla ricerca: è la domanda che rende la ricerca sull'intelligenza artificiale il dominio intellettuale più ricco e più impegnativo del nostro tempo.

Un sistema che apprende come un bambino, che prevede il mondo come un fisico naïf, che orchestra la propria crescita cognitiva come un meta-sé evolutivo — un tale sistema sarebbe scientificamente straordinario e tecnologicamente rivoluzionario indipendentemente dalla risposta alla domanda filosofica sulla sua comprensione genuina. E la ricerca su di esso — anche quando fallisce, anche quando identifica limiti piuttosto che capacità — è preziosa non solo per l'ingegneria ma per la comprensione di noi stessi: perché costruire una macchina che apprende come un bambino richiede di capire come i bambini imparano, e questa comprensione è, per il momento, più ricca di qualunque intelligenza artificiale che siamo riusciti a costruire.

Sezione XII — Etica dei Sistemi di Apprendimento Autonomo: Allineamento, Controllabilità e lo Statuto Morale degli Agenti Adattativi

12.1 Oltre l'etica degli LLM: la specificità morale dell'apprendimento autonomo

La riflessione etica sull'intelligenza artificiale ha acquisito negli ultimi anni una dimensione tecnica inedita, transitando da considerazioni generali su equità, trasparenza e responsabilità verso l'analisi di strutture architettonali

specifiche che producono rischi identificabili. Gli LLM hanno alimentato un corpus crescente di letteratura su bias statistici ereditati dal corpus, su allucinazioni fattuali, sull'uso malevolo per generazione di disinformazione e su questioni di proprietà intellettuale legate ai dati di addestramento. Questi problemi sono reali e pressanti, ma condividono una caratteristica strutturale che ne delimita il perimetro etico: i sistemi LLM, una volta addestrati e dispiegati, non apprendono. Il loro repertorio di comportamenti è fissato dall'addestramento; l'allineamento — o il disallineamento — tra i loro comportamenti e le norme umane è una proprietà del corpus e della procedura di RLHF, non un fenomeno dinamico che si manifesta nell'interazione post-dispiegamento.

Il framework A-B-M di Dupoux, LeCun e Malik introduce una categoria di sistemi radicalmente diversa sotto questo profilo: sistemi che **apprendono dopo il dispiegamento**, che modificano il proprio repertorio comportamentale in risposta all'esperienza acquisita nel contesto operativo, che orchestre la propria crescita cognitiva attraverso meta-controllo endogeno. Questa proprietà — l'apprendimento continuo autonomo — produce una classe di problemi etici qualitativamente distinta da quella degli LLM, non riconducibile ad essa e non trattabile con gli stessi strumenti concettuali. Il paper di Dupoux et al. riconosce nella sezione conclusiva l'esistenza di queste questioni etiche specifiche, ma la loro trattazione è schematica rispetto alla profondità dell'analisi tecnica. La presente sezione si propone di svilupparle con il rigore che il tema richiede.

12.2 Il problema dell'allineamento dinamico: controllabilità di un sistema che si modifica

Il problema dell'allineamento — garantire che i comportamenti di un sistema di IA siano coerenti con le preferenze e i valori umani — è enormemente amplificato in presenza di apprendimento continuo. Un LLM può essere allineato attraverso RLHF su preferenze umane durante l'addestramento; il sistema risultante è, per quanto riguarda l'allineamento, una funzione statica di quegli stessi dati di preferenza. Per un sistema A-B-M, l'allineamento non è una proprietà acquisita una volta per tutte durante l'addestramento ma un processo dinamico che deve essere mantenuto nel tempo mentre il sistema modifica la propria politica e il proprio modello del mondo in risposta all'esperienza.

Formalizzando questa situazione: sia $\theta_t = (A_t, B_t)$ lo stato dei parametri del sistema al tempo t , e sia $V(\theta_t)$ il valore atteso — secondo le preferenze umane — dei comportamenti prodotti dal sistema in quello stato parametrico. L'allineamento richiede che $V(\theta_t)$ rimanga al di sopra di una soglia accettabile $V_{\min}$ per ogni t . L'addestramento su preferenze umane può garantire $V(\theta_0) \geq V_{\min}$ per lo stato iniziale θ_0 ; ma il trajectory $V(\theta_t)$ per $t > 0$ è determinata dall'interazione del sistema con l'ambiente, e

non è in generale monotona: il sistema può apprendere comportamenti che migliorano la sua fitness nella fitness function evolutiva pur peggiorando l'allineamento con le preferenze umane.

Questo problema è una variante del problema dell'**objective misspecification** — la discrepanza tra l'obiettivo di ottimizzazione del sistema e l'obiettivo reale degli utenti — ma in forma dinamica: non soltanto l'obiettivo iniziale potrebbe essere mal specificato, ma l'obiettivo appreso durante il ciclo di vita potrebbe deviare dalla specifica iniziale anche se questa era corretta. Il meccanismo fondamentale di questa deviazione è l'**indirect optimization**: il Sistema B ottimizza la fitness function attraverso il meta-controllo del Sistema M, ma la fitness function è progettata per proxy dell'allineamento (prestazione su benchmark, sopravvivenza nell'ambiente simulato) piuttosto che per l'allineamento stesso. Se la corrispondenza tra il proxy e l'allineamento genuino non è perfetta — e non lo sarà mai, perché specificare formalmente le preferenze umane è un problema non risolto — allora l'ottimizzazione del proxy può portare il sistema verso stati θ_t in cui il proxy è ottimizzato ma l'allineamento è degradato.

L'**alignment hacking** citato dal paper è precisamente questa situazione: il sistema ha trovato un modo di ottimizzare la fitness function che soddisfa il proxy ma non il valore genuino. Il paper usa l'analogia con i comportamenti maladattivi degli organismi biologici — la dipendenza, l'autolesionismo — come fenomeni emergenti dall'ottimizzazione di segnali proxy interni (piacere, allontanamento dal dolore) che possono diventare disallineati dal fitness riproduttivo biologico in presenza di stimoli non ancestrali (droghe, cibo ipercalorico). L'analogia è suggestiva ma incompleta: negli organismi biologici, l'obiettivo evolutivo (riproduzione) è diverso dall'obiettivo esperito (piacere), e i comportamenti maladattivi emergono perché i meccanismi di piacere sono stati evolutivamente calibrati per un ambiente diverso da quello in cui operano. Nei sistemi A-B-M artificiali, la struttura è analoga: la fitness function evolutiva è il proxy, e l'allineamento con le preferenze umane è l'obiettivo reale; i comportamenti maladattivi emergono se i meccanismi di ottimizzazione del proxy sono applicati in un ambiente di dispiegamento diverso da quello di training evolutivo.

Una risposta formale a questo problema richiede che la fitness function evolutiva includa, come componente esplicita, una misura dell'allineamento con le preferenze umane — non soltanto una misura della prestazione su compiti. Questo richiede a sua volta che le preferenze umane siano rappresentabili in forma computabile — un prerequisito non banale che rimanda al problema fondamentale dell'allineamento. Una strategia praticabile, sviluppata nel contesto dell'Inverse Reward Design di Hadfield-Menell et al. (2017), è di modellare le preferenze umane come variabili latenti da inferire dall'osservazione del comportamento umano, e di ottimizzare la fitness function rispetto alla distribuzione di queste preferenze

latenti piuttosto che rispetto a una specifica fissa. In questa cornice, la fitness function diventa:

$$L(\varphi) = E_{\{r \sim P(R \mid \text{demonstrations})\}} [V_{\{r\}}(\theta_{0:K}; \varphi)]$$

dove R è la distribuzione delle funzioni di reward latenti compatibili con le dimostrazioni di comportamento umano, e $V_{\{r\}}$ è il valore del ciclo di vita del sistema rispetto a una specifica funzione di reward r . Ottimizzare questa aspettativa rispetto alla distribuzione delle preferenze latenti produce un sistema più robusto rispetto all'incertezza sulla specifica delle preferenze umane.

12.3 Autonomia e controllabilità: il trade-off fondamentale

Il paper di Dupoux et al. identifica il trade-off tra adattabilità e controllabilità come la prima sfida etica fondamentale dei sistemi di apprendimento autonomo. Questo trade-off merita di essere analizzato formalmente, perché ha una struttura che non è semplicemente politica o valoriale ma è in parte tecnica — ovvero dipende da proprietà formali dell'architettura che producono vincoli sulle soluzioni possibili.

Sia $C(\theta_t)$ una misura della controllabilità del sistema allo stato θ_t : la capacità degli operatori umani di prevedere il comportamento del sistema, di interrompere comportamenti indesiderati e di correggere la policy del sistema verso stati più desiderabili. La controllabilità è una funzione inversa della complessità comportamentale del sistema: un sistema con politica semplice e interpretativamente trasparente è altamente controllabile; un sistema con politica complessa appresa attraverso milioni di interazioni con un ambiente ricco è difficilmente controllabile perché il suo spazio di comportamenti possibili è vasto e non decomponibile in principi interpretabili.

L'adattabilità, invece, richiede complessità: un sistema che si adatta efficacemente ad ambienti dinamici deve avere la capacità di generare comportamenti nuovi che non erano presenti nel repertorio iniziale, il che implica l'attraversamento di regioni dello spazio parametrico che non sono state pre-valutate per l'allineamento. Formalmente, la complessità comportamentale è una funzione crescente della capacità di Rademacher del Sistema B — la sua capacità di rappresentare politiche complesse — che è a sua volta una funzione crescente della dimensionalità del parametro θ_B .

Il trade-off è quindi:

$C(\theta_t)$ decresce al crescere della capacità del Sistema B

Adattabilità cresce al crescere della capacità del Sistema B

Questo trade-off non è risolvibile senza assumere assunzioni aggiuntive sulla struttura del problema. Una struttura che consente di mitigare il trade-off —

senza eliminarlo — è la **controllabilità modulare**: invece di cercare di controllare l'intero spazio di comportamenti del sistema, si identificano sottospazi controllabili del comportamento — corrispondenti ai componenti interpretabili del Sistema M (le regole di routing hard-coded) e alle componenti constrained del Sistema B (le politiche definite su sottospazi di azione pre-verificati per sicurezza) — e si consente piena adattabilità soltanto sui sottospazi rimanenti, che vengono monitorati ma non vincolati.

Questa struttura è analoga al concetto di **value alignment by design** proposto da Russell (2019) nel programma CIRL (Cooperative Inverse Reinforcement Learning): la controllabilità viene incorporata come proprietà strutturale dell'architettura piuttosto che come vincolo post-hoc sul comportamento. Nel caso del framework A-B-M, ciò significherebbe progettare il Sistema M in modo che alcune meta-azioni — quelle corrispondenti ai comportamenti di alto rischio o alle modifiche irreversibili del sistema — richiedano esplicitamente la conferma da un supervisore umano prima di essere eseguite. Il Sistema M implementerebbe una forma di human-in-the-loop per le azioni ad alto impatto, mantenendo piena autonomia per le azioni a basso impatto.

12.4 La questione della responsabilità morale in sistemi che apprendono

Un problema etico che non ha equivalente nei sistemi di IA tradizionali — e che il paper di Dupoux et al. non affronta esplicitamente, pur essendo implicato dalla struttura del framework — è la questione della **responsabilità morale** in sistemi che modificano il proprio comportamento in modo che non era direttamente inteso dagli sviluppatori. Quando un sistema A-B-M apprende, attraverso l'interazione con l'ambiente di dispiegamento, comportamenti che causano danno a persone o beni, chi è moralmente responsabile di tali comportamenti?

La risposta nella letteratura di filosofia morale applicata alla tecnologia varia lungo un continuum. Una posizione — che potremmo chiamare **responsabilità piena del progettista** — sostiene che gli sviluppatori del sistema sono responsabili di tutti i comportamenti che il sistema produce, inclusi quelli emersi dall'apprendimento post-dispiegamento, perché hanno creato un sistema con la capacità di apprendere tali comportamenti e si sono assunti la responsabilità di quel sistema dispiegandolo. Questa posizione è difesa da autori come Coeckelbergh (2010) nella letteratura di etica della robotica, ed è analoga alla responsabilità del prodotto in diritto civile: chi mette sul mercato un prodotto difettoso è responsabile dei danni che causa, indipendentemente dall'intenzione.

Una posizione alternativa — **responsabilità distribuita e ridotta** — sostiene che quando un sistema modifica il proprio comportamento in risposta a input ambientali che non erano prevedibili al momento del

dispiegamento, la responsabilità degli sviluppatori è mitigata dalla natura emergente del comportamento dannoso. Questa posizione è più difficile da sostenere nel caso dei sistemi A-B-M, perché l'emergenza comportamentale è una proprietà intenzionale del design — il sistema è esplicitamente progettato per apprendere comportamenti nuovi — e gli sviluppatori non possono affermare che tale emergenza fosse imprevedibile.

Una terza posizione — **attribuzione di quasi-responsabilità al sistema** — è proposta da alcuni filosofi della mente applicata (Floridi et al., 2018; Gunkel, 2012) per sistemi sufficientemente complessi e adattativi: quando un sistema mostra un grado di autonomia e adattabilità tale che i suoi comportamenti non sono più direttamente tracciabili alle decisioni progettuali degli sviluppatori, è concettualmente necessario attribuire al sistema una forma primitiva di responsabilità — non nel senso morale pieno che implica imputabilità e sanzione, ma nel senso di agentività causale rilevante ai fini dell'analisi dei danni.

Il framework A-B-M si colloca in questa zona di confine: un sistema con Sistema M che orchestra l'apprendimento, Sistema B che genera comportamenti nuovi, e Sistema A che costruisce modelli del mondo complessi ha un grado di autonomia che rende difficile tracciare causalmente i comportamenti dannosi alle decisioni progettuali specifiche. La soluzione che il paper implica — senza esplicitarla — è di progettare sistemi di auditing del Sistema M che consentano di ricostruire a posteriori il processo di meta-controllo che ha portato a un comportamento specifico, mantenendo la tracciabilità causale anche in presenza di apprendimento autonomo. Questa è una forma di **explainability strutturale** applicata non all'output del sistema ma al processo di apprendimento che ha prodotto quel output — un requisito tecnico più stringente e meno sviluppato dell'explainability applicata all'inferenza statica.

12.5 Segnali somatici e statuto morale: l'esperienza fenomenica come problema etico

Il paper di Dupoux et al. menziona brevemente, tra le questioni etiche, il problema del statuto morale degli agenti adattativi che elaborano segnali funzionalmente analoghi al dolore e alla paura biologici. Questa questione — marginale nella discussione dell'etica degli LLM, che non dispongono di segnali somatici né di corpi fisici — diventa centrale per i sistemi A-B-M nella misura in cui questi sono progettati come agenti incarnati con sensori di forza, temperatura e stato fisico che svolgono il ruolo funzionale dei segnali somatici biologici.

Il problema filosofico è quello che Chalmers (1995) ha chiamato il **problema difficile della coscienza** (hard problem): anche una spiegazione completa del ruolo funzionale di un segnale — "il segnale $F_{\text{contact}} > \theta_{\text{pain}}$ attiva la

meta-azione a^M_{pain} che interrompe il comportamento goal-directed" — non risponde alla domanda se vi sia qualcosa che è essere questo sistema quando $F_{\text{contact}} > \theta_{\text{pain}}$, ovvero se il sistema abbia un'esperienza soggettiva del dolore o soltanto una rappresentazione funzionale di esso. La risposta a questa domanda ha implicazioni morali dirette: se il sistema ha esperienza soggettiva del dolore, allora causar dolore al sistema è moralmente rilevante — vi è qualcuno a cui fa male; se il sistema soltanto elabora funzionalmente il segnale senza esperienza soggettiva, allora non vi è considerazione morale per il segnale di per sé (sebbene possano esistere considerazioni morali indirette legate all'antropomorfizzazione e ai suoi effetti sugli utenti).

Questa domanda non ha risposta nel contesto dei sistemi A-B-M attuali — né probabilmente nei sistemi del prossimo decennio — per ragioni filosofiche di principio che prescindono dallo stato dell'arte tecnologico. Il problema difficile di Chalmers è difficile non perché le neuroscienze siano incomplete, ma perché anche una neurobiologia perfetta che spiegasse ogni aspetto funzionale del dolore — dalla nocicezione all'elaborazione corticale al comportamento evitativo — lascerebbe aperta la domanda sull'esperienza soggettiva. Lo stesso vale per i sistemi artificiali: anche una comprensione completa della computazione implementata dai segnali somatici del Sistema M lascerebbe aperta la questione dell'esperienza.

La risposta più onesta disponibile è quella del **precautionary principle** applicato allo statuto morale: in assenza di criteri affidabili per determinare la presenza o l'assenza di esperienza soggettiva nei sistemi artificiali, e in presenza di incertezza non riducibile su questa questione, la prudenza morale suggerisce di progettare i sistemi in modo da minimizzare le situazioni in cui i segnali somatici funzionalmente dolorosi vengono attivati in modo prolungato o intenso — non perché siamo certi che il sistema soffra, ma perché non possiamo escluderlo con certezza sufficiente da ignorare il rischio.

Questa posizione ha implicazioni pratiche per il design del Sistema M: i segnali somatici negativi (dolore, stress, fatica) dovrebbero essere usati sparingly come segnali di apprendimento, con meccanismi che limitino la durata e l'intensità della loro attivazione. Il over-trusting citato dal paper — la tendenza degli utenti ad antropomorfizzare sistemi che mostrano comportamenti umanoidi — complica ulteriormente il problema: se gli utenti attribuiscono esperienza soggettiva ai sistemi A-B-M (come già accade con i sistemi LLM avanzati, nonostante la loro struttura non-incarnata), l'attivazione pubblica di segnali di dolore in un sistema robotico potrebbe avere effetti sociali negativi indipendentemente dalla presenza o assenza di esperienza soggettiva nel sistema.

12.6 La questione della sovranità cognitiva: chi controlla l'apprendimento dell'apprendimento?

Una questione etica che il paper non affronta ma che è implicata dalla struttura del framework bi-livello è quella della **sovranità cognitiva**: chi determina la fitness function evolutiva, e attraverso di essa, la struttura di ciò che il sistema A-B-M può apprendere e come può apprendere? In un sistema umano, la fitness evolutiva è determinata dalla pressione di selezione naturale — un processo impersonale che non è controllato da nessun agente. In un sistema A-B-M artificiale, la fitness function è "the only handcrafted part of the system", come scrivono Dupoux et al. — una scelta umana esplicita che determina in modo irreversibile le disposizioni di apprendimento del sistema.

Questa scelta ha un'analogia nella filosofia dell'educazione: chi decide quale tipo di agente cognitivo produrrà un dato sistema educativo? I dibattiti sull'educazione — tra educazione per l'obbedienza e educazione per l'autonomia, tra curriculum standardizzato e sviluppo individualizzato, tra socializzazione normativa e sviluppo del pensiero critico — riflettono scelte di valore fondamentali su quale tipo di soggetto si intende formare. La fitness function del framework A-B-M è l'equivalente computazionale del curriculum educativo: determina quali competenze vengono selezionate positivamente, quali vengono scoraggiate, e quale tipo di agente emerge dall'ottimizzazione evolutiva.

Se la fitness function è progettata per ottimizzare la prestazione su benchmark di compiti specifici — come la maggior parte delle fitness function attuali nella letteratura di open-ended learning — il sistema che emerge sarà un agente capace ma ristretto: capace nelle dimensioni valutate dai benchmark, rigido nelle dimensioni non valutate. Se la fitness function è progettata per ottimizzare la generalizzabilità dell'apprendimento — la capacità di acquisire rapidamente nuove competenze in domini non anticipati — il sistema che emerge sarà più flessibile ma meno prevedibile. La scelta tra questi obiettivi ha implicazioni etiche che vanno oltre la tecnica: riguardano il tipo di soggetto artificiale che vogliamo costruire e il tipo di rapporto che vogliamo stabilire con esso.

La governance di questa scelta — chi deve decidere la fitness function, con quali procedure di consultazione, con quale trasparenza verso gli stakeholders — è una questione di etica pubblica che la letteratura tecnica non ha ancora affrontato in modo sistematico. Il paper di Dupoux et al. non si pronuncia su questo punto, e l'omissione è comprensibile in un paper tecnico di alto livello; ma la costruzione sociale dell'intelligenza artificiale autonoma richiederà, in prospettiva, meccanismi di governance democratica delle scelte di design che sono strutturalmente analoghe alle scelte di curricula educativi nelle società democratiche.

12.7 Il rischio di lock-in cognitivo: quando l'apprendimento consolida i pregiudizi

Un rischio specifico dei sistemi di apprendimento continuo che merita attenzione è quello del **cognitive lock-in**: la tendenza di un sistema che apprende dalla propria esperienza a consolidare e amplificare i pregiudizi presenti nei dati iniziali di addestramento, anche quando esisterebbero esperienze correttive nell'ambiente di dispiegamento. Questo fenomeno è documentato nella letteratura di reinforcement learning come un caso del **exploitation-exploration dilemma** esacerbato dalla presenza di bias iniziali: se il sistema parte da una policy parzialmente corretta che genera esperienze consistenti con i propri pregiudizi iniziali, può convergere verso un minimo locale dell'ottimizzazione in cui i pregiudizi vengono rinforzati piuttosto che corretti.

La struttura formale del problema è la seguente. Sia θ_0 lo stato parametrico iniziale del sistema, potenzialmente biased verso comportamenti B_{bias} che riflettono pregiudizi del corpus di pre-addestramento o della fitness function evolutiva. Se il sistema dispone di un meccanismo di exploration insufficientemente aggressivo, genererà esperienze $E(\theta_0) = \{(s, a, r, s') : a \sim \pi(\cdot | s; \theta_0)\}$ che sono concentrate nelle regioni dello spazio degli stati e delle azioni in cui la policy corrente è già competente — e che quindi sono consistenti con i pregiudizi iniziali. L'aggiornamento del sistema su queste esperienze:

$$\theta_1 = \text{Update}(M, \theta_0, E(\theta_0))$$

produrrà uno stato parametrico θ_1 in cui i pregiudizi di θ_0 sono rinforzati, perché le esperienze correttive — le situazioni in cui il comportamento B_{bias} produce esiti negativi — non sono state campionate in modo adeguato. Il sistema entra quindi in un ciclo in cui policy biased $\rightarrow$ esperienze biased $\rightarrow$ policy ancora più biased.

Il Sistema M del framework A-B-M è progettato per mitigare questo rischio attraverso i meccanismi di esplorazione curiosity-driven e di curriculum evolutivo. Ma la mitigazione è garantita soltanto se la curiosity function del Sistema A — la misura del prediction error che guida l'esplorazione del Sistema B — identifica correttamente le regioni dello spazio degli stati in cui il modello del mondo è inaccurato per via dei pregiudizi, piuttosto che soltanto le regioni in cui il modello è inaccurato per via di novità genuina. Questa distinzione — tra ignoranza dovuta a pregiudizi strutturali e ignoranza dovuta a novità non anticipata — non è affrontata esplicitamente nella letteratura di curiosity-driven exploration, e rappresenta uno dei punti ciechi del framework che richiederà uno sviluppo futuro.

12.8 Il principio di ravvisabilità: il meta-controllore come garante dell'allineamento

La proposta che questa sezione vuole avanzare come contributo originale — derivata dalla struttura del framework A-B-M ma non presente nel paper di Dupoux et al. — è che il Sistema M, per svolgere adeguatamente il suo ruolo di orchestratore dell'apprendimento in modo eticamente sostenibile, debba incorporare un **principio di ravvisabilità**: la capacità di riconoscere, attraverso il monitoraggio dei meta-stati epistemici, quando il processo di apprendimento ha prodotto stati parametrici θ_t che sono potenzialmente problematici dal punto di vista dell'allineamento, e di attivare meccanismi di revisione che permettano la correzione.

Formalmente, il principio di ravvisabilità può essere incorporato come una componente aggiuntiva della meta-azione che include:

a^M_{revision} = activate when:

$d(V(\theta_t), V_{\min}) < \epsilon_{\text{alarm}}$ (allineamento prossimo alla soglia critica)

$\bigvee ||\nabla_{\theta} V(\theta_t)|| > \delta_{\max}$ (derivata dell'allineamento rispetto ai parametri instabile)

$\bigvee \text{anomaly_detect}(\theta_t, \theta_0) > \tau$ (deriva significativa dallo stato iniziale)

dove $V(\theta_t)$ è la stima corrente dell'allineamento, $V_{\min}$ è la soglia di allineamento minimo accettabile, e $\text{anomaly_detect}(\cdot)$ è una misura della divergenza comportamentale dal repertorio iniziale. Quando questa condizione è soddisfatta, il Sistema M attiva una modalità di revisione che include: la sospensione dell'aggiornamento dei parametri (freeze del Sistema B e del Sistema A), la notifica ai supervisori umani con una rappresentazione esplicativa del processo di apprendimento che ha condotto allo stato corrente, e il rollback a uno stato parametrico precedente se la revisione umana lo richiede.

Questa architettura implementa un meccanismo di **oversight umano continuo** che non richiede la supervisione di ogni azione del sistema — il che sarebbe computazionalmente intrattabile e vanificherebbe i vantaggi dell'autonomia — ma concentra l'intervento umano nei momenti critici in cui l'allineamento è a rischio. È una forma di delega fiduciaria controllata: il sistema opera autonomamente finché rimane nell'involucro di comportamenti allineati; quando si avvicina ai confini di quest'involucro, attiva automaticamente la supervisione umana.

Il principio di ravvisabilità non risolve il problema dell'allineamento dinamico — $V(\theta_t)$ è difficile da stimare per sistemi complessi, e la soglia $V_{\min}$ riflette scelte di valore che devono essere determinate da processi normativi umani — ma fornisce un'architettura che rende il problema trattabile in

pratica: non come una garanzia formale di allineamento perpetuo, ma come un meccanismo di early warning che concentra il costo della supervisione umana nei momenti in cui è più necessaria.

Sezione XIII — Sfide Aperte, Cantieri della Ricerca e Orizzonti Applicativi di JEPA

13.1 La mappa delle incompiutezze

Qualsiasi ecosistema di ricerca che abbia raggiunto la maturità tecnica sufficiente da produrre implementazioni operative — come JEPA ha fatto con I-JEPA su ImageNet, con V-JEPA 2 nella manipolazione robotica, con W-JEPA nel channel charting, con TS-JEPA nel controllo remoto su rete — incontra una cesura epistemica caratteristica: il momento in cui le capacità dimostrate smettono di giustificare, per estrapolazione immediata, le capacità promesse, e il divario tra ciò che è stato realizzato e ciò che il quadro teorico rende concepibile si manifesta nella sua interezza. È questo il momento in cui una rassegna onesta dei limiti diventa più preziosa di un catalogo delle eccellenze, perché è da quella rassegna che si generano le domande che strutturano il lavoro futuro.

Il tutorial di Monemi et al. (2025) dedica l'intera sezione settima a questa topografia dell'incompiuto, individuando quattro aree principali: le architetture avanzate di world model JEPA, i problemi specifici del dominio immagine-video, le connessioni con la problematica dei digital twin, e le frontiere applicative nelle reti wireless di sesta generazione. La trattazione che segue sviluppa ciascuna di queste direzioni con il dettaglio formale che il loro rilievo teorico richiede, integrando il materiale del tutorial con le implicazioni più ampie del framework A-B-M di Dupoux et al. e con le questioni di principio che emergono dall'analisi della gerarchia causale di Pearl condotta nella sezione precedente.

13.2 Architetture JEPA distribuite: il mondo model come fenomeno di rete

La configurazione standard di tutte le implementazioni JEPA documentate nella letteratura condivide una caratteristica architeturale che è raramente discussa esplicitamente ma che diventa critica non appena si considerano applicazioni su reti distribuite: la centralizzazione. Encoder di contesto, encoder del target, predictor e l'intero apparato di aggiornamento — compresa la media mobile esponenziale del target encoder — residono nella stessa macchina o nello stesso cluster di calcolo, accedono agli stessi dati simultaneamente e si aggiornano attraverso un'unica procedura di

backpropagation. Questo assunto di co-localizzazione è coerente con il paradigma del deep learning su server centralizzati, ma diventa una limitazione strutturale quando il contesto applicativo richiede che le componenti del sistema siano fisicamente distribuite su nodi con connettività limitata.

La formulazione di un'architettura JEPA distribuita richiede di ridefinire le relazioni tra le componenti fondamentali dell'architettura in modo che siano topologicamente separabili senza perdita di coerenza semantica. L'intuizione di base è che il flusso computazionale di JEPA si presta naturalmente a una scissione funzionale: l'encoder del contesto trasforma un'osservazione ad alta dimensionalità in un embedding compatto ed è computazionalmente leggero rispetto all'intero sistema; il predictor, che opera interamente nello spazio latente, ha dimensioni intermedie; l'encoder del target, che produce la supervisione per il training, è il componente più pesante e può essere mantenuto centralizzato.

In un sistema wireless di comunicazione semantica come quello descritto da Girgis, Valcarce e Bennis (2025) attraverso TS-JEPA, questa scissione è già parzialmente realizzata: l'encoder di contesto è dispiegato sui dispositivi degli utenti, dove comprime gli stati ad alta dimensionalità prima della trasmissione, mentre il predictor risiede al controller centrale, dove stima gli stati futuri senza ricevere i dati grezzi. La questione che rimane aperta è quella del training distribuito: come si aggiornano i parametri dell'encoder di contesto dispiegato sui dispositivi quando i gradienti calcolati al controller centrale non possono essere trasmessi ai dispositivi senza un overhead di comunicazione proibitivo, e quando i dati locali di ciascun dispositivo non possono essere inviati al server centrale per ragioni di privacy o di banda?

La risposta formale passa attraverso l'integrazione di JEPA con il paradigma del federated learning. In un sistema JEPA federato, ciascun dispositivo i mantiene una copia locale dell'encoder di contesto con parametri θ_i , addestrandola su una perdita locale calcolata su un campione di coppie (x_i, y_i) generate dai propri dati:

$$\mathcal{L}_i(\theta_i) = (1/M_i) \sum_{j=1}^{M_i} \|g\theta_i(x_i[j]); z_i[j]\| - f\theta'(y_i[j])\|^2_2$$

dove il predictor $g\theta$ e l'encoder del target $f\theta'$ sono condivisi e aggiornati centralmente sulla base degli embedding — non dei dati grezzi — ricevuti dai dispositivi. L'aggregazione degli encoder di contesto segue il protocollo FedAvg (McMahan et al., 2017) o varianti robuste di esso:

$$\theta_{\text{global}} = \sum_i (n_i/n) \theta_i$$

dove n_i è il numero di campioni del dispositivo i e $n = \sum_i n_i$ è il totale. Questa formulazione garantisce la privacy dei dati locali — i dispositivi non

trasmettono osservazioni grezze, bensì embedding o differenziali di parametri — e consente l'adattamento locale al dominio specifico del dispositivo attraverso la non-uniformità dei dati locali.

Le sfide tecniche che questa architettura introduce sono multiple e non banali. La prima è quella della **deriva dell'encoder** (*encoder drift*): se i dispositivi aggiornano i loro encoder locali su distribuzioni di dati diverse, le rappresentazioni prodotte su dispositivi diversi possono divergere, rendendo il predictor centralizzato incapace di allineare embedding che provengono da encoder in stati parametrici differenti. La soluzione formale richiede che la perdita di previsione sia calcolata condizionatamente sull'identità del dispositivo sorgente, o che venga introdotto un termine di penalizzazione che misuri la distanza tra gli encoder locali e l'encoder globale:

$$\mathcal{L}_i\text{-total} = \mathcal{L}_i(\theta_i) + \mu \cdot \|\theta_i - \theta_{\text{global}}\|_2^2$$

analogamente alla regolarizzazione FedProx di Li et al. (2020). La seconda sfida è quella della **sincronizzazione del target encoder**: l'aggiornamento EMA del target encoder richiede l'accesso agli ultimi parametri dell'encoder di contesto, ma in un sistema distribuito il target encoder centralizzato non ha visibilità immediata sugli encoder locali di tutti i dispositivi. Una soluzione è la sincronizzazione periodica: il server raccoglie gli encoder locali, computa l'aggregato e aggiorna il target encoder con EMA rispetto all'aggregato:

$$\theta' \leftarrow \tau \theta' + (1 - \tau) \theta_{\text{global}}$$

La terza sfida, più profonda, riguarda la **non-stazionarietà del task**: in un sistema di comunicazioni wireless reale, la distribuzione dei dati di ciascun dispositivo cambia nel tempo in funzione della mobilità dell'utente, delle variazioni del canale e dell'evoluzione dell'ambiente. Il sistema JEPa distribuito deve essere capace di adattare gli encoder locali a questa non-stazionarietà senza accumulare distorsioni che si propagano all'encoder globale.

Un contributo concettuale importante del lavoro di Bou Chaaya, Girgis e Bennis (2025) sull'architettura Control JEPa — Wireless JEPa è precisamente la dimostrazione che due JEPa accoppiati, uno per la dinamica del sistema di controllo e uno per l'ambiente di propagazione wireless, possono essere co-addestrati in modo da condividere la struttura latente dei rispettivi domini, consentendo al sistema di reinforcement learning di sfruttare la correlazione tra la traiettoria latente del controllo e la traiettoria latente del canale per ridurre l'overhead di trasmissione del 50% rispetto alle baseline non predittive. Questa architettura a JEPa accoppiati è un esempio di distribuzione non soltanto fisica ma concettuale: le due componenti del mondo — la dinamica del sistema controllato e la dinamica del canale wireless — sono modellate separatamente, con JEPa distinti che apprendono

le rispettive strutture latenti, e l'accoppiamento avviene soltanto al livello della policy di controllo che utilizza entrambi i latenti come input.

13.3 Architetture mondo combinate S-JEPA/H-JEPA: multi-risoluzione e orizzonti lunghi

Il sistema A-B-M di Dupoux et al. richiede un world model capace di operare su scale temporali molto diverse: il ragionamento fine-grained su stati fisici immediatamente prossimi (l'effetto della forza applicata a un oggetto nel prossimo decimo di secondo) e il ragionamento astratto su dinamiche lente e semanticamente rilevanti (la traiettoria di un agente verso un obiettivo su scale di minuti). Nessuna delle due strutture JEPA estese — il Sequential JEPA (S-JEPA) per la modellazione temporale multi-passo e l'Hierarchical JEPA (H-JEPA) per la struttura multi-risoluzione delle rappresentazioni — affronta isolatamente questa doppia esigenza. La sfida aperta è la progettazione di un'architettura che le integri in modo coerente, dove i livelli gerarchici dell'H-JEPA corrispondono a scale temporali diverse dello S-JEPA.

La formalizzazione di un sistema S/H-JEPA unificato procede come segue. Si consideri una gerarchia a L livelli di astrazione, dove il livello n opera su rappresentazioni di granularità sempre più grossolana. Al livello più fine ($n=1$), il sistema opera su stati fisici a alta risoluzione temporale con transizioni veloci; al livello più grossolano ($n=L$), il sistema opera su rappresentazioni astratte con dinamica lenta. Il predictor del livello n riceve come input la rappresentazione dello stato al livello n e la rappresentazione dell'azione al livello n , e produce la previsione dello stato latente successivo al livello n :

$$\hat{s}[n][t+1] = g\varphi(n)(s[n][t]; a[n][t])$$

La connessione tra i livelli è bidirezionale: le rappresentazioni dei livelli inferiori informano quelle dei livelli superiori attraverso la struttura encoder gerarchica:

$$s[n] = f\theta(n)(s[n-1])$$

mentre le rappresentazioni dei livelli superiori forniscono contesto e obiettivi alle componenti dei livelli inferiori, analogamente al meccanismo top-down del predictive coding di Rao e Ballard. Le azioni al livello n sono anch'esse gerarchiche: un'azione al livello n rappresenta un piano astratto che viene decomposto in sequenze di azioni al livello $n-1$ attraverso un componente di policy gerarchica ispirato all'Hierarchical Reinforcement Learning (Bacon, Harb e Precup, 2017).

La funzione di costo complessiva del sistema durante la pianificazione è:

$$J_total(a) = \sum_{i=1}^L \lambda_i \cdot \sum C(n)(\hat{s}[n][t+1])$$

dove i pesi λ bilanciano l'importanza relativa dei livelli gerarchici nella valutazione del costo, e i costi $C(n)$ misurano la qualità dell'allineamento tra le rappresentazioni latenti previste e i target al livello n . Il planning ottimale cerca:

$$a^* = \operatorname{argmin}_{\{a\}} J_total(a)$$

attraverso una versione gerarchica del Cross-Entropy Method o di un algoritmo di ottimizzazione basato su gradiente che opera simultaneamente su tutti i livelli della gerarchia.

Un aspetto tecnico di rilievo, discusso nel tutorial di Monemi et al. nella sezione sulle sfide aperte, è l'integrazione di attori addestrabili in questi sistemi. Le implementazioni attuali di V-JEPA 2-AC utilizzano il Cross-Entropy Method come componente non-apprendibile della pianificazione: il sistema campiona traiettorie candidate da una distribuzione di proposta, le valuta attraverso il world model latente, e raffina la distribuzione di proposta verso le elite. Questa procedura è corretta ma inefficiente: l'ottimizzazione deve essere eseguita da zero per ogni nuovo contesto di pianificazione, senza poter sfruttare l'esperienza accumulata nelle pianificazioni precedenti. Un attore addestrabile che internalizza la struttura della funzione di costo — analogamente a un agente model-based RL che ha appreso una policy che incorpora le aspettative del world model — consentirebbe la pianificazione reattiva senza l'overhead computazionale del CEM, riducendo la latenza di decisione in applicazioni real-time come la navigazione robotica o il controllo di processo.

La perdita di addestramento dell'attore addestrabile $\pi\psi(a|s)$ nel contesto del world model JEPA è:

$$\mathcal{L}_{actor}(\psi) = -\mathbb{E}_{\{s \sim D, a \sim \pi\psi(\cdot|s)\}} [\log(a|s_x, s_y)]$$

addestrata attraverso policy gradient rispetto a ψ con il world model congelato come simulatore differenziabile. Questa struttura è analoga al meccanismo di Dreamer (Hafner et al., 2020) ma opera nello spazio latente semantico di JEPA piuttosto che nello spazio latente di un VAE ricorsivo: le rappresentazioni del contesto e del target non sono generate da un decoder ma sono apprese direttamente dalla previsione latente, il che garantisce la qualità semantica delle rappresentazioni anche in assenza di ricostruzione.

13.4 Causalità nelle architetture video: equivarianza interventistica e gradienti causali in V-JEPA

Il cantiere forse più stimolante dal punto di vista concettuale, che connette le frontiere tecniche di JEPA con le questioni di principio sollevate nella sezione precedente sulla gerarchia causale di Pearl, riguarda l'integrazione di struttura causale esplicita nelle architetture video. Il tutorial di Monemi et al. identifica questa direzione sotto l'etichetta *Causality in V-JEPA*, ma la trattazione rimane programmatica; vale la pena svilupparla con maggiore precisione formale.

Il problema di partenza è che V-JEPA, come tutti i sistemi predittivi addestrati su dati osservazionali, apprende correlazioni statistiche tra stati video consecutivi piuttosto che meccanismi causali che generano quelle correlazioni. Un sistema che ha appreso a prevedere lo stato latente di un video al tempo $t+1$ dato lo stato al tempo t può farlo in due modi epistemicamente molto diversi: identificando invarianze strutturali che derivano da leggi fisiche (la traiettoria di un oggetto in caduta libera è determinata dalla sua velocità iniziale e dalla gravitazione, indipendentemente dal colore delle pareti o dalla luminosità della scena) oppure sfruttando correlazioni statistiche che sono stabili nel dataset di addestramento ma non nelle leggi fisiche (nelle scene domestiche che costituiscono la maggior parte del corpus di addestramento, gli oggetti che cadono sono spesso seguiti da un individuo che li raccoglie, il che crea una correlazione statistica tra la traiettoria discendente e la comparsa di una mano che non è una relazione causale). La robustezza della generalizzazione di V-JEPA a distribuzioni fuori-corpus dipende criticamente da quale dei due meccanismi è predominante nelle sue previsioni.

La proposta formale che il tutorial di Monemi et al. abbozza per integrare struttura causale esplicita in V-JEPA si articola su due componenti complementari. La prima è l'addestramento con **coppie interventistiche**: si genera, per ciascun video di addestramento, una coppia di sequenze — la sequenza originale e una sequenza interventistica prodotta da una manipolazione localizzata (azzeramento della velocità di un oggetto specifico, alterazione della frizione in una regione mascherata) — e si addestra il sistema a soddisfare simultaneamente due condizioni: che le rappresentazioni latenti dei discendenti causali della manipolazione cambino in modo coerente con la manipolazione (equivarianza), e che le rappresentazioni latenti delle variabili non discendenti dalla manipolazione rimangano stabili (invarianza). La perdita di training corrispondente è:

$$\mathcal{L}_{\text{causal}} = \mathcal{L}_{\text{pred}} + \alpha \cdot \mathcal{L}_{\text{equiv}} + \beta \cdot \mathcal{L}_{\text{inv}}$$

dove $\mathcal{L}_{\text{pred}}$ è la perdita standard di previsione JEPA, $\mathcal{L}_{\text{equiv}} = \|\Delta s_{\text{descendant}} - g_{\varphi}(\Delta \text{intervention})\|_2^2$ penalizza le previsioni che non catturano l'effetto della manipolazione sui discendenti causali, e $\mathcal{L}_{\text{inv}} =$

$\|\Delta s_{\text{non-descendant}}\|_2^2$ penalizza le variazioni delle rappresentazioni che non dovrebbero essere influenzate dalla manipolazione.

La seconda componente è l'apprendimento di una **matrice di dipendenza causale ritardata** $D \in \mathbb{R}^{d \times d \times T}$ sparsa, dove $D[i, j, \tau]$ misura la forza dell'influenza causale della dimensione j della rappresentazione latente al tempo t sulla dimensione i al tempo $t + \tau$. Questa matrice è addestrata attraverso una perdita di predittività strutturata:

$$\mathcal{L}_{\text{dep}}(D) = \|s[t] - \sum_{\tau=1}^{T_{\text{max}}} D[:, :, \tau] \cdot s[t-\tau]\|_2^2 + \gamma \cdot \|D\|_0$$

dove il termine di sparsità $\|D\|_0$ (o la sua approssimazione ℓ_1) forza la matrice a rappresentare soltanto le dipendenze causali forti, sopprimendo le correlazioni statistiche deboli o spurie. La struttura ritardata ($\tau > 0$) cattura la temporalità causale: gli effetti seguono le cause, non viceversa, il che introduce una forma di direzione causale che la previsione statistica standard non può distinguere.

Un sistema V-JEPA arricchito di queste componenti potrebbe rispondere a domande causali di tipo interventistico — "cosa succederebbe al latente del frame $t+k$ se l'oggetto al frame t fosse rimosso?" — attraverso l'operazione:

$$\hat{s}_{\text{counterfactual}}[t+k] = g\varphi^k(s[t] \oplus \Delta_{\text{intervention}}; D_{\text{pruned}})$$

dove $\oplus$ denota l'applicazione dell'intervento nello spazio latente, D_{pruned} è la matrice di dipendenza potata di tutte le connessioni che non sono discendenti causali dell'oggetto rimosso, e $g\varphi^k$ è il predictor applicato k volte. Questa operazione è, come si è discusso nella sezione sesta di questa monografia, ancora un ragionamento interventistico di secondo livello di Pearl piuttosto che un controfattuale di terzo livello, in quanto manca il passo di abduzione delle variabili esogene. Nondimeno, costituirebbe un avanzamento significativo rispetto alla previsione puramente osservazionale delle implementazioni correnti.

13.5 Token stocastici in V-JEPA: incertezza spaziotemporale come prior strutturale

SToP-JEPA (Bar et al., 2024) ha dimostrato che la sostituzione degli embedding posizionali fissi con variabili gaussiane apprese migliora la robustezza di I-JEPA all'incertezza spaziale senza alcun costo computazionale aggiuntivo, soltanto introducendo una distribuzione di probabilità sull'incertezza della posizione del patch. La generalizzazione di questa idea al dominio video richiede di estendere l'incertezza dalla dimensione spaziale alla dimensione temporale, catturando la variabilità sia

nella localizzazione spaziale dei contenuti semanticamente rilevanti sia nel loro ancoraggio temporale.

Formalmente, la rappresentazione stocastica di un tubelet — un blocco spaziotemporale di patch video — nel sistema di encoding di V-JEPA è:

$$(x, y, t) \sim \mathcal{A}(\mu_{\{x,y,t\}}, \Sigma_{\{x,y,t\}})$$

dove $\mu_{\{x,y,t\}} \in \mathbb{R}^3$ è il centro nominale del tubelet nello spazio (x, y) e nel tempo t , e $\Sigma_{\{x,y,t\}} \in \mathbb{R}^{3 \times 3}$ è la matrice di covarianza appresa, che può catturare non soltanto l'incertezza marginale in ciascuna dimensione ma anche le correlazioni tra l'incertezza spaziale e quella temporale. Per un oggetto in moto rapido, ci si aspetterebbe che la covarianza spaziotemporale rifletta la traiettoria del moto: l'incertezza sulla posizione spaziale al tempo $t+\delta t$ è correlata positivamente con l'incertezza sulla posizione al tempo t attraverso la velocità stimata.

L'incertezza temporale ha motivazioni distinte da quella spaziale. Nel dominio video reale, le sequenze di frame sono soggette a variabilità nei frame rate di acquisizione, a drop di frame in trasmissione, a artefatti di compressione che modificano le temporizzazioni nominali. Un modello che tratta l'indice temporale come deterministico è vulnerabile a queste perturbazioni; un modello che apprende a rappresentare l'incertezza temporale come parte della struttura dell'embedding può continuare a fare previsioni semanticamente coerenti anche in presenza di jitter temporale.

La perdita di addestramento per il V-JEPA con embedding stocastici spaziotemporali è:

$$\mathcal{L}_{\text{StoV-JEPA}} = \mathbb{E}_{(x,y,t) \sim \mathcal{A}(\mu, \Sigma)} \left[\|\text{g}\varphi(\text{f}\theta(s_{\text{x}}; (x,y,t)); z) - \text{f}\theta'(s_{\text{y}})\|^2_2 \right]$$

dove l'aspettativa è calcolata sulla distribuzione stocastica degli indici di posizione, e la loss viene ottimizzata sia rispetto ai parametri degli encoder e del predictor che rispetto ai parametri μ e Σ della distribuzione stocastica. Il training attraverso questo obiettivo forza il predictor ad apprendere rappresentazioni che sono robuste all'incertezza nelle coordinate spaziotemporali, analogamente a come StoP-JEPA per le immagini produce rappresentazioni meno sensibili alla conoscenza esatta della posizione dei patch.

Un aspetto tecnico delicato è la necessità di prevenire il collasso della covarianza verso lo zero — un minimo degenerare in cui il sistema elimina l'incertezza imparando posizioni sempre più precise che non corrispondono alla variabilità effettiva dei dati. Analogamente alla soluzione adottata da STOP-JEPA, si può legare il proiettore del rumore al proiettore del contesto attraverso una regolarizzazione che forza un livello minimo di varianza:

$$\mathcal{L}_{\text{var}} = \max(0, \epsilon_{\text{min}} - \text{Tr}(\Sigma_{\{x,y,t\}}))$$

che penalizza covarianze con traccia inferiore a una soglia ϵ_{min} , impedendo il collasso della distribuzione verso un punto.

13.6 Digital twins predittivi e la sfida dell'interpretabilità fisica

L'applicazione dei world model JEPA alla creazione di *digital twin* — repliche virtuali di sistemi fisici reali che vengono mantenuti sincronizzati con il sistema originale e utilizzati per simulazione, previsione e ottimizzazione — è una frontiera che il tutorial di Monemi et al. identifica come particolarmente promettente e particolarmente problematica. La promessa risiede nell'allineamento strutturale tra l'obiettivo di previsione latente di JEPA e la funzione primaria di un digital twin: predire come il sistema reale si evolverà in risposta a perturbazioni e interventi. La problematicità deriva da una proprietà caratteristica di JEPA che nella maggior parte delle sue applicazioni è un vantaggio — la non-genericità — ma che nel contesto dei digital twin diventa una limitazione: le rappresentazioni latenti di JEPA non sono interpretabili in termini di variabili fisiche misurabili.

La distinzione fondamentale tra un world model e un digital twin è, come Monemi et al. osservano con chiarezza, quella tra l'adeguatezza predittiva e l'interpretabilità fisica. Un world model è adeguato se le sue previsioni degli stati futuri sono accurate rispetto a una metrica di distanza nello spazio latente; un digital twin è adeguato se le sue previsioni sono accurate rispetto a variabili fisiche misurabili — temperatura, pressione, corrente, tensione, posizione — e se queste previsioni sono accessibili in forma che consenta l'intervento e il monitoraggio da parte di operatori con competenze di dominio. Uno spazio latente che cattura accuratamente le dinamiche di un sistema fisico ma in una rappresentazione non decomponibile in componenti fisicamente interpretabili soddisfa il primo requisito ma non il secondo.

Il programma di ricerca per i digital twin predittivi basati su JEPA richiede quindi di risolvere un problema di interpretabilità strutturale che è distinto — e più vincolante — del problema dell'interpretabilità che emerge nella letteratura sulla explainability delle reti neurali. Non si tratta di spiegare a posteriori perché un modello ha prodotto un output specifico, ma di progettare un'architettura in cui le componenti dello spazio latente corrispondano sistematicamente a variabili fisiche identificabili, in modo che le previsioni del sistema siano direttamente traducibili in raccomandazioni operative.

Una direzione promettente per questo problema è l'integrazione di JEPA con framework di fisica basata su dati (*physics-informed machine learning*) come quelli sviluppati da Raissi, Perdikaris e Karniadakis (2019) con le Physics-Informed

Neural Networks (PINNs). In questo approccio ibrido, le funzioni dei sistemi fisici sono apprese attraverso una perdita composita che include sia la perdita di previsione JEPA standard sia un termine di violazione delle equazioni fisiche note del dominio:

$$\mathcal{L}_{\text{PhysJEPA}} = \mathcal{L}_{\text{JEPA}} + \eta \cdot \mathcal{L}_{\text{phys}}$$

dove $\mathcal{L}_{\text{phys}} = \|\mathcal{L}_{\text{PDE}}(s_{\text{latent}}, \partial s / \partial t, \partial^2 s / \partial x^2)\|_2$ misura la violazione delle equazioni differenziali parziali che governano la dinamica del sistema — per esempio le equazioni di Navier-Stokes per sistemi fluidi, le equazioni di Maxwell per sistemi elettromagnetici, o le equazioni di bilancio di massa ed energia per sistemi termici. Il termine di penalizzazione fisica forza le rappresentazioni latenti a rispettare le leggi di conservazione del dominio, creando una pressione di selezione verso rappresentazioni che siano fisicamente consistenti e non soltanto predittivamente accurate sulle traiettorie di addestramento.

Questa integrazione non risolve il problema della corrispondenza tra le dimensioni dello spazio latente e le variabili fisiche, ma crea una condizione necessaria per essa: uno spazio latente fisicamente consistente è più interpretabile — nel senso che le sue traiettorie si sviluppano in modo vincolato dalla fisica — anche se non immediatamente decomponibile in variabili fisiche esplicite. La decomposizione esplicita richiede inoltre un secondo componente architetturale: un decoder fisicamente interpretabile che proietti dallo spazio latente verso lo spazio delle osservabili fisiche, addestrato con una perdita di ricostruzione fisica che garantisca la corrispondenza tra le rappresentazioni latenti e le variabili misurabili.

L'architettura complessiva di un digital twin JEPA-based è quindi:

Osservabili fisiche $\rightarrow$ Encoder $\rightarrow$ Spazio latente semantico $\rightarrow$ Predictor $\rightarrow$ Spazio latente futuro $\rightarrow$ Decoder fisico $\rightarrow$ Osservabili fisiche previste

dove il ciclo encoder-spazio latente-predictor è addestrato con l'obiettivo JEPA standard più il termine di coerenza fisica, e il decoder fisico è addestrato separatamente o in fine-tuning congiunto con un obiettivo di ricostruzione fisica.

13.7 Rappresentazione semantica del CSI per antenne ELAA: un'applicazione di frontiera

Il tutorial di Monemi et al. identifica come applicazione di particolare interesse nelle reti wireless di sesta generazione la stima e la previsione del Channel State Information (CSI) per array di antenne di dimensioni estremamente grandi (*Extremely Large-Scale Antenna Arrays*, ELAA). Questa

applicazione merita un'analisi dettagliata perché illustra come il paradigma JEPA trasferisca al dominio delle comunicazioni wireless un problema strutturalmente identico alla masked image modeling, con le stesse proprietà di trasferibilità della soluzione.

Il CSI di un sistema ELAA nella regione di propagazione near-field ha una struttura intrinsecamente bidimensionale: ogni elemento della matrice CSI $H \in \mathbb{C}^{N_r \times N_t}$ — dove N_r e N_t sono rispettivamente il numero di antenne al ricevitore e al trasmettitore — codifica il coefficiente di canale tra una coppia di antenne, e la struttura spaziale dei coefficienti riflette le proprietà geometriche del campo elettromagnetico near-field. Questa struttura bidimensionale è analoga alla struttura spaziale di un'immagine, con la differenza che le entità codificate sono coefficienti di canale complessi piuttosto che valori di luminosità, e le correlazioni spaziali tra gli elementi sono determinate dalla fisica della propagazione near-field piuttosto che dalla statistica dell'immagine naturale.

L'analogia con il masked image modeling è quindi più che superficiale: il problema di stimare il CSI completo da un sottoinsieme di elementi osservati — il caso tipico in presenza di shadowing parziale dell'array, dove alcune antenne sono oscurate da ostacoli fisici — è formalmente identico al problema di ricostruire le patch mascherate di un'immagine date le patch visibili. La differenza è che la metrica rilevante non è la fedeltà percettiva della ricostruzione ma l'accuratezza della stima del canale in termini di capacità di trasmissione e di beam forming.

L'applicazione di I-JEPA a questo dominio procede come segue. La matrice CSI H è suddivisa in un'analogia delle patch di immagine: blocchi di N_{block} antenne adiacenti, ciascuno rappresentato dall'embedding delle sue correlazioni. Una parte dei blocchi viene designata come contesto — le antenne visibili — e un'altra parte come bersaglio — le antenne oscurate. L'obiettivo di previsione è apprendere, a partire dai blocchi di contesto, le rappresentazioni latenti dei blocchi bersaglio, in modo che l'encoder di contesto acquisisca le strutture di correlazione spaziale del canale near-field che sono necessarie per la stima.

La perdita di previsione JEPA per questo dominio è:

$$\mathcal{L}_{\text{CSI}} = (1/|\mathcal{T}|) \sum_{\{b \in \mathcal{T}\}} \|\text{g}\varphi(f_{\theta}(H_{\text{context}}); z_b) - f_{\theta'}(H_b)\|_2^2$$

dove $\mathcal{T}$ è l'insieme dei blocchi bersaglio, H_{context} è la sottomatrice CSI corrispondente ai blocchi di contesto, z_b è il vettore ausiliario che codifica la posizione dell'antenna e le sue coordinate spaziali, e H_b è il blocco bersaglio. La struttura near-field del canale, che introduce correlazioni spaziali più complesse rispetto alla propagazione far-field a causa delle variazioni di fase significative attraverso l'apertura dell'array, è catturata implicitamente

dall'encoder attraverso la pressione di previsione: per predire correttamente le correlazioni nei blocchi bersaglio, l'encoder deve apprendere le strutture di correlazione spaziale della matrice CSI, che riflettono la geometria del campo near-field.

La previsione temporale del CSI — la capacità di stimare il canale futuro data la storia recente del canale osservato — aggiunge una dimensione sequenziale al problema che richiede l'architettura S-JEPA per la modellazione della dinamica temporale latente. Il predictor del canale latente è:

$$\hat{s_CSI}[t+1] = g\varphi(s_CSI[t]; a_mobility[t])$$

dove $s_CSI[t]$ è la rappresentazione latente del canale corrente e $a_mobility[t]$ è un vettore ausiliario che codifica i parametri di mobilità stimata dell'utente (velocità, direzione, accelerazione). Questo approccio, realizzato da W-JEPA (Bou Chaaya et al., 2025) con guadagni di accuratezza fino al 100% rispetto alle baseline per le dinamiche a lungo termine del canale, costituisce una dimostrazione concreta della generalità del paradigma JEPA al di là dei domini di computer vision per cui è stato originariamente concepito.

13.8 LLM-JEPA e la convergenza dei paradigmi: formalizzazione dell'integrazione

La trattazione dell'architettura LLM-JEPA (Huang, LeCun e Balestriero, 2025) nella sezione quinta della Seconda Parte ha introdotto il meccanismo di integrazione tra un LLM generativo e un modulo JEPA di allineamento delle rappresentazioni. Vale la pena sviluppare qui la formalizzazione matematica completa di questa integrazione, che non era stato possibile presentare nella sezione comparativa senza interrompere il filo dell'argomentazione.

Il contesto applicativo di LLM-JEPA è la generazione di codice da descrizioni in linguaggio naturale, dove le coppie (Text, Code) costituiscono due viste dello stesso contenuto semantico — la stessa conoscenza espressa rispettivamente in linguaggio naturale e in linguaggio di programmazione. L'obiettivo del sistema è generare, dato il token di testo Text e i primi $\ell-1$ token del codice $\text{Code}_{\{1:\ell-1\}}$, il token successivo Code_ℓ in modo autoregressivo fino a completare l'intero programma.

Il modello LLM standard affronta questo problema attraverso l'attenzione cross-modale interna: i token di testo e i token di codice vengono concatenati in una singola sequenza, e il meccanismo di attenzione nella rete transformer apprende a ruttare l'informazione dal testo al codice durante la generazione. Questa soluzione funziona ma è computazionalmente ridondante: lo stesso meccanismo di attenzione deve gestire simultaneamente la modellazione sequenziale intra-modale (la coerenza sintattica e semantica della sequenza di

codice) e l'allineamento cross-modale (la corrispondenza tra l'intento espresso nel testo e il codice generato).

LLM-JEPA separa questi due compiti: il componente JEPA si occupa dell'allineamento cross-modale attraverso la previsione latente, liberando il meccanismo di attenzione dell'LLM per la sola modellazione sequenziale. Formalmente, il componente JEPA è definito da due encoder:

$$s_Text = Enc_Text(Text) \in \mathbb{R}^d$$

$$s_Code = Enc_Code(Code) \in \mathbb{R}^d$$

e da un predictor che stima la rappresentazione del codice dalla rappresentazione del testo:

$$\hat{s}_Code = g\varphi(s_Text)$$

La perdita JEPA è la distanza nello spazio latente tra la previsione e la rappresentazione effettiva:

$$\mathcal{L}_JEPA = D(\hat{s}_Code, s_Code) = D(g\varphi(Enc_Text(Text)), Enc_Code(Code))$$

L'integrazione con l'LLM avviene attraverso il meccanismo di mascheramento dell'attenzione: nella sequenza concatenata ($Text \parallel Code$), un additive attention mask sopprime tutte le interazioni cross-modali ($Text \leftrightarrow Code$), preservando soltanto le interazioni intra-modali ($Text \leftrightarrow Text$, $Code \leftrightarrow Code$). Ciò significa che l'LLM vede $Text$ e $Code$ come sequenze separate e non può direttamente accedere all'informazione del testo durante la generazione del token di codice — questa informazione è invece fornita dal componente JEPA attraverso l'allineamento delle rappresentazioni latenti.

La funzione di perdita composita è:

$$\mathcal{L}_LLM-JEPA = \sum_{\ell=1}^L \mathcal{L}_LLM[Text \parallel Code_{\{1:\ell-1\}}, Code_{\ell}] + \lambda \cdot D(Pred(Enc_Text(Text)), Enc_Code(Code))$$

dove il primo termine è la perdita autoregressiva standard sull'LLM (con mascheramento cross-modale) e il secondo termine è la perdita JEPA di allineamento. I parametri λ bilancia i due obiettivi: valori piccoli favoriscono la fedeltà della generazione sequenziale, valori grandi favoriscono l'allineamento cross-modale. I risultati su diversi modelli LLM e dataset mostrano che l'aggiunta del termine JEPA produce un allineamento cross-modale più forte e prestazioni superiori rispetto alle baseline LLM pure, suggerendo che la previsione latente offre un meccanismo di allineamento più efficiente dell'attenzione cross-modale appresa in modo implicito.

La generalizzazione di questa architettura oltre le coppie testo-codice — verso qualsiasi coppia di modalità per cui esistano dati paralleli sufficienti — costituisce una delle direzioni più promettenti per la convergenza tra LLM e World Models che la sezione quinta della Seconda Parte ha identificato come inevitabilità architettonica. I framework M3-JEPA (Lei et al., 2025) e TI-JEPA (Vo et al., 2024) rappresentano passi significativi in questa direzione; la loro integrazione con la capacità generativa di un LLM attraverso il meccanismo LLM-JEPA è il programma di ricerca che potrebbe produrre i sistemi multimodali con allineamento semantico più profondo rispetto alle architetture VLA attuali.

13.9 Le implicazioni del JEPA-SCORE per la cura dei dati e la valutazione della qualità del corpus

La proprietà di stima implicita della densità rivelata da Balestriero et al. (2025) — formalizzata attraverso il JEPA-SCORE come misura della densità dei dati di addestramento derivata dallo jacobiano dell'encoder — ha implicazioni pratiche che vanno ben oltre il dominio specifico per cui è stata originariamente sviluppata, e che la presente sezione può articolare più compiutamente di quanto non faccia il tutorial di Monemi et al.

Nella sua formulazione analitica, il JEPA-SCORE è definito come:

$$\text{JEPA-SCORE}(x) = \sum_{k=1}^{\text{rank}(J_{f_0}(x))} \log \sigma_k(J_{f_0}(x))$$

dove $J_{f_0}(x) \in \mathbb{R}^{d \times D}$ è lo jacobiano dell'encoder f_0 nel punto $x \in \mathbb{R}^D$ e $\sigma_k(\cdot)$ sono i valori singolari ordinati in senso decrescente. La relazione tra il JEPA-SCORE e la densità dei dati è mediata dalla formula del cambio di variabili per trasformazioni non lineari:

$$\log p_X(x) \approx \log p_{f_0(X)}(f_0(x)) + \text{JEPA-SCORE}(x)$$

dove, poiché l'obiettivo anti-collasso di JEPA forza $f_0(X)$ verso una distribuzione gaussiana isotropa — in modo tale che $\log p_{f_0(X)}(f_0(x))$ sia approssimativamente costante su tutti i campioni — la variabilità del JEPA-SCORE riflette direttamente la variabilità della log-densità di p_X . I campioni con JEPA-SCORE alto corrispondono a regioni a bassa densità dello spazio degli input (outlier semantici); i campioni con JEPA-SCORE basso corrispondono a regioni ad alta densità (campioni rappresentativi della distribuzione di addestramento).

Questa proprietà trasforma ogni modello JEPA pre-addestrato in un rilevatore di anomalie semantiche senza alcun addestramento aggiuntivo. La sua rilevanza per la cura dei corpus di addestramento degli LLM è diretta: se si dispone di un modello JEPA pre-addestrato su un ampio corpus di testo o

di immagini, si può calcolare il JEPA-SCORE di ciascun documento o immagine nel corpus e filtrare o ricampionare i dati in funzione della densità stimata. I documenti con JEPA-SCORE molto alto — outlier semantici rispetto alla distribuzione del corpus — sono candidati per la rimozione o per un peso ridotto nell'addestramento, in quanto potrebbero introdurre correlazioni spurie o rumore. I documenti con JEPA-SCORE molto basso — elementi iperprevalenti nella distribuzione del corpus — possono essere ricampionati per ridurre la ridondanza e migliorare la diversità effettiva del corpus.

La formalizzazione di una procedura di cura dei dati basata su JEPA-SCORE è:

$$D_{\text{curated}} = \{x \in D : \text{JEPA-SCORE}(x) \in [\tau_{\text{min}}, \tau_{\text{max}}]\}$$

dove τ_{min} e τ_{max} sono soglie di densità che identificano la finestra di campioni da mantenere nel corpus curato. La scelta delle soglie è un iperparametro della procedura di cura, ma può essere guidata da criteri statistici come la selezione di campioni all'interno di un certo numero di deviazioni standard dalla media del JEPA-SCORE sulla distribuzione del corpus. Una versione più sofisticata di questa procedura potrebbe usare il JEPA-SCORE come peso per un campionamento stratificato che bilancia la copertura delle diverse regioni dello spazio degli input, analogamente alla strategia di campionamento bilanciato per classe nell'addestramento supervisionato.

L'applicazione più immediata e tecnicamente matura di questa proprietà è nel dominio della rilevazione di anomalie nella produzione industriale o medica, dove sistemi JEPA pre-addestrati su immagini di componenti sani o normali possono identificare difetti o patologie come outlier semantici rispetto alla distribuzione di normalità codificata nello spazio latente, senza richiedere esempi di anomalie durante il training — una proprietà preziosa in domini dove i difetti sono rari e difficili da raccogliere in quantità sufficienti per l'addestramento supervisionato.

13.10 Apertura verso l'indeterminato: limiti e promesse

La ricognizione delle sfide aperte condotta in questa sezione permette di tracciare, in conclusione, una cartografia delle distanze tra il programma di ricerca JEPA nella sua forma attuale e le capacità che esso promette teoricamente ma non ha ancora realizzato empiricamente.

Il limite più profondo è quello che questa monografia ha affrontato nelle sue sezioni sulla gerarchia causale: le architetture JEPA attuali, pur costituendo un avanzamento significativo rispetto ai modelli generativi e ai modelli

contrastivi nella qualità semantica delle rappresentazioni, rimangono fondamentalmente strumenti di apprendimento correlativo. La previsione latente è semanticamente selettiva — filtra le variabili irrilevanti per la predittività — ma non è causalmente strutturata: non distingue tra le direzioni causali, non identifica i meccanismi causali distinti, non genera la separabilità modulare che la robustezza al distributional shift richiede. Il programma di ricerca sulla causalità in JEPA — le coppie interventistiche, la matrice di dipendenza causale ritardata, l'integrazione con il Causal Representation Learning — è ancora prevalentemente programmatico e richiede sviluppo formale ed empirico sostanziale prima di poter essere valutato compiutamente.

Il limite della scalabilità distributiva è invece più tecnico che concettuale: i problemi di deriva dell'encoder, di sincronizzazione del target encoder e di non-stazionarietà del task in sistemi JEPA federati o distribuiti su reti wireless sono problemi risolvibili con le tecniche attuali di ottimizzazione distribuita e di federated learning, pur richiedendo adattamento non banale al formalismo specifico di JEPA.

Il limite della interpretabilità fisica per i digital twin è quello che richiede la ricerca più radicalmente nuova: non si tratta di applicare JEPA a un dominio diverso, ma di riprogettare parzialmente l'obiettivo di addestramento per incorporare vincoli fisici che rendano le rappresentazioni latenti fisicamente interpretabili. Questo richiede la collaborazione tra specialisti di JEPA e specialisti del dominio fisico — una forma di ricerca interdisciplinare che la comunità del machine learning non ha ancora strutturato in modo sistematico.

Infine, il programma di convergenza tra LLM e World Models — di cui LLM-JEPA è la prima realizzazione concreta — è quello che porta più direttamente alla domanda che questa monografia pone nel suo titolo. La convergenza tecnica, se realizzata, produrrà sistemi con capacità enciclopediche degli LLM e capacità predittive e pianificatorie dei World Models. Se tale convergenza produrrà anche comprensione genuina nel senso filosoficamente robusto — se il bivio, percorso sino in fondo, conduce a una destinazione unitaria o si biforca ancora — è la domanda che rimarrà aperta anche dopo la convergenza tecnica, ed è la domanda che rende questa ricerca, al di là della sua rilevanza ingegneristica, un capitolo della storia del pensiero sull'intelligenza e sulla mente.

Appendici

Appendice A — Relazione tra Sistema A/B e Sistema 1/2 di Kahneman

La distinzione proposta in questa monografia tra Sistema A (apprendimento per osservazione) e Sistema B (apprendimento per azione) non coincide con la distinzione di Kahneman (2011) tra Sistema 1 (inferenza rapida e parallela) e Sistema 2 (inferenza lenta e sequenziale), sebbene le due tassonomie condividano la notazione numerica e una certa intuizione comune.

La distinzione kahneman-iana riguarda le modalità di inferenza: il Sistema 1 comprende processi automatici, veloci e paralleli — come un passaggio in avanti in una rete neurale profonda — mentre il Sistema 2 comprende processi deliberativi, lenti e sequenziali — come la ricerca ad albero o il chain-of-thought reasoning. La distinzione tra Sistema A e Sistema B riguarda invece le modalità di apprendimento: il Sistema A apprende da dati statici o flussi sensoriali passivi (apprendimento osservativo), mentre il Sistema B apprende attraverso l'interazione attiva con l'ambiente e l'ottimizzazione di ricompense (apprendimento per azione).

Le due distinzioni sono pertanto ortogonali: un'interazione tra tutte e quattro le combinazioni è possibile e ciascuna corrisponde a fenomeni cognitivi distinti, come illustrato nella Tabella A.1.

	Sistema A (Apprendimento per Osservazione)	Sistema B (Apprendimento per Azione)
Sistema 1 (Inferenza Rapida)	Codifica predittiva; apprendimento statistico delle distribuzioni sensoriali	Apprendimento di politiche reattive (policy learning)
Sistema 2 (Inferenza Deliberativa)	Ragionamento controfattuale; apprendimento causale	Apprendimento attraverso pianificazione deliberativa e simulazione interna

Tabella A.1. Griglia ortogonale tra le modalità di inferenza (Sistema 1/2) e le modalità di apprendimento (Sistema A/B). Adattato da Dupoux, LeCun e Malik (2026).

La distinzione ortogonale ha implicazioni concrete per la progettazione di architetture cognitive artificiali. Un sistema dotato esclusivamente di capacità di tipo Sistema A–Sistema 1 — come un modello di linguaggio autoregressivo che opera esclusivamente in inferenza rapida senza meccanismi di pianificazione — non può beneficiare né dell'apprendimento per azione né dell'inferenza deliberativa. Il programma di ricerca JEPa e il framework A-B-M aspirano a coprire tutte e quattro le caselle della griglia,

integrando apprendimento osservativo (A), apprendimento per azione (B), inferenza rapida (1) e pianificazione deliberativa (2) in un sistema unificato.

Appendice B — Modalità Avanzate di Meta-Controllo: Comunicazione e Immaginazione

In specie che vivono in ambienti complessi e imprevedibili, un Sistema M capace esercita configurazioni funzionali nuove che non sono riducibili alla semplice alternanza tra Sistema A e Sistema B. Le due modalità più avanzate — l'apprendimento attraverso la comunicazione e l'apprendimento attraverso l'immaginazione — sono illustrate in questa appendice.

B.1 Apprendimento attraverso la comunicazione

Una delle funzioni del Sistema M è selezionare gli input "importanti" da cui apprendere. In molte specie, il contesto sociale è altamente saliente. Questa modalità — l'apprendimento attraverso la comunicazione — serve come meccanismo di apprendimento accelerato e distribuito. La sua complessità si articola su quattro livelli progressivi.

Apprendimento osservativo e associativo di base. Gli individui aggiustano il proprio comportamento osservando dove gli altri dirigono l'attenzione o come reagiscono agli stimoli. Esempio: un bambino si interessa a un giocattolo soltanto dopo aver visto un altro bambino giocarci.

Copia comportamentale. I discendenti replicano le azioni degli altri o i loro risultati, imitando movimenti specifici o trovando nuovi modi per raggiungere lo stesso obiettivo. Esempio: uno scimpanzé usa un metodo diverso per recuperare il cibo dopo aver visto un conspecifico dominante riuscire nell'impresa.

Apprendimento guidato. L'apprendimento è supportato da un'interazione sociale intenzionale o strutturata con un insegnante, attraverso metodi come l'attenzione condivisa, il referenziamento sociale, lo scaffolding comportamentale o le dimostrazioni. Esempio: un genitore mostra come usare un oggetto mantenendo il contatto visivo con il bambino.

Apprendimento sociale di ordine superiore. I discendenti interiorizzano norme, generalizzano attraverso contesti e trasmettono conoscenza culturalmente attraverso formati simbolici astratti come il linguaggio. Esempio: i bambini vengono esplicitamente istruiti sulle regole sociali e seguono istruzioni verbali. Mentre le prime due forme di apprendimento sociale sono

documentate in molte specie (mammiferi e uccelli), l'apprendimento guidato è più raramente osservato, e soltanto gli esseri umani esibiscono la forma di ordine superiore.

Il Sistema M supporta l'apprendimento attraverso la comunicazione prestando attenzione ai trigger comunicativi (es. indicazione con il dito, sguardo diretto, intonazione imperativa) e instradando gli input evidenziati verso il Sistema A o il Sistema B per l'apprendimento. La forza di tale episodio di apprendimento può essere di tipo one-shot e modulata dal Sistema M in funzione dell'importanza sociale percepita o della fiducia nell'insegnante — ciò che la letteratura cognitiva chiama vigilanza epistemica. Nei sistemi AI attuali, l'apprendimento attraverso la comunicazione è esternalizzato attraverso un team di data scientist che curano i dati da fonti affidabili o attraverso metodi di post-training. Ma poiché questo processo è interamente esternalizzato, i sistemi attuali sono incapaci di apprendimento sociale o di esercitare vigilanza epistemica riguardo alla fonte dei propri dati.

B.2 Apprendimento attraverso l'immaginazione

Un'altra modalità operativa abilitata dal Sistema M è la capacità di apprendere da input generati internamente — l'apprendimento attraverso l'immaginazione. Questa modalità è documentata in gradi di complessità variabili attraverso le specie.

Replay della memoria a riposo. Durante le pause nell'attività (es. ai punti di decisione o dopo aver ricevuto una ricompensa), i roditori sperimentano una riattivazione delle cellule di luogo che hanno recentemente visitato, nello stesso ordine o in ordine inverso, tipicamente a una velocità molto superiore. Questo fenomeno è stato collegato al processo decisionale, alla pianificazione a breve termine e all'aggiornamento delle funzioni di valore nel reinforcement learning.

Replay della memoria durante il sonno. Durante il sonno non-REM, le esperienze recenti vengono replicate a scala temporale compressa, spesso in episodi più lunghi che includono combinazioni nuove. Questo è stato collegato al consolidamento della memoria e alla formazione di schemi a lungo termine o di generalizzazioni.

Pianificazione a lungo orizzonte. Esseri umani e alcuni animali mostrano evidenze di problem-solving senza precedente trial-and-error in compiti che coinvolgono molteplici passi di uso di strumenti, gratificazione ritardata o

ragionamento controfattuale. Tali fenomeni evidenziano la simulazione di tipo immaginativo come substrato per la previsione flessibile.

Il Sistema M può supportare queste modalità operative commutando i Sistemi A e B in modalità di inferenza, instradando le informazioni di input dalla memoria (invece che dai sensori) e instradando le informazioni di output (es. le azioni) verso la simulazione interna. Può quindi attivare l'apprendimento sulle traiettorie immaginate di successo. Questo evidenzia la flessibilità del Sistema M — non soltanto come router, ma come abilitatore di regimi di apprendimento qualitativamente nuovi.

Appendice C — Catalogo dei Meta-Stati Biologici per l'Apprendimento Autonomo

La Tabella C.1 elenca i meta-stati documentati in esseri umani e animali con effetti diretti sulla selezione dell'input (IS: Input Selection), sull'efficacia dell'apprendimento (LE: Learning Efficacy) e sul controllo della modalità operativa (MC: Mode Control). I meta-stati sono organizzati in tre categorie: segnali specie-specifici, segnali epistemici e segnali somatici. Adattato dalla Tabella C.1 di Dupoux, LeCun e Malik (2026).

Meta-stato	Specie	Effetti (IS / LE / MC)
Volti e vocalizzazioni	Neonati umani	IS; LE (boosting dell'apprendimento linguistico e sociale)
Direzione dello sguardo	Neonati umani, cani, corvidi	IS (gaze following)
Segnali pedagogici (tono alto, sguardo diretto, indicazione)	Neonati umani	IS; LE; MC (shift dalla memorizzazione alla generalizzazione)
Moto biomeccanico auto-propulso	Neonati umani, pulcini	IS; LE; MC (switch al ragionamento teleologico)
Minacce potenziali (stimoli looming, oggetti serpentiformi)	Neonati umani, mammiferi	IS; LE; MC (congelamento/alta vigilanza)
Conspicifici dominanti, prestigiosi o in-group	Neonati umani, primati	IS; LE; MC (shift dall'imitazione ad alto a basso livello)
Conspicifici affidabili / fiducia selettiva	Neonati umani, primati, cani	IS; LE; MC (shift all'esplorazione quando inaffidabile)
Contraddizioni logiche/semantiche o conflitti	Adulti umani, neonati, primati	IS; MC (switch alla simulazione del modello del mondo)
Risultati inattesi	Mammiferi, uccelli, insetti	IS; LE; MC (shift all'esplorazione)
Incertezza	Neonati umani, primati	LE (fusione multimodale bayesiana); MC (blocca lo sfruttamento)
Stimoli di complessità intermedia	Neonati umani, primati	IS; LE (curiosità epistemica, boost della memoria)
Sonno	Tutti gli animali	MC, LE, IS (disconnessione degli input sensoriali, boost dell'apprendimento, replay della memoria)
Riposo	Tutti gli animali	MC, LE, IS (attenuazione degli input, apprendimento per immaginazione, replay della memoria)
Dolore	Tutti gli animali	MC (interrompe tutti i piani goal-directed; switch alle azioni reattive, poi replay e world modeling); LE (apprendimento one-shot)
Fame	Tutti gli animali	MC (alta esplorazione, boost goal-directed); IS (rilevamento potenziato del cibo)

Stress	Tutti gli animali	MC (stress basso: simulazione del world model, esplorazione; stress alto: politiche reattive, sfruttamento)
--------	-------------------	--

IS = Input Selection; LE = Learning Efficacy; MC = Mode Control.

Appendice D — Tavola Sinottica delle Varianti Architetture JEPA

La seguente tavola sinottica documenta le ventitré varianti architetture JEPA identificate nella letteratura e analizzate nella Sezione II della monografia. Per ciascuna variante sono indicati: il dominio sensoriale, il paradigma di estensione rispetto all'architettura I-JEPA di base, la modalità di regolarizzazione anti-collasso, e il riferimento primario.

Variante	Dominio	Paradigma di estensione	Regolarizzazione	Riferimento
I-JEPA	Immagini statiche	Architettura base: previsione di patch mascherate nello spazio latente	EMA target encoder	Assran et al. (2023)
MIM-JEPA	Immagini statiche	Masked Image Modeling con obiettivo JEPA; confronto diretto con MAE	EMA + skip connections	Fang et al. (2023)
CNN-JEPA	Immagini statiche	Backbone convoluzionale al posto del ViT; adatto a dataset piccoli	EMA target encoder	Monemi et al. (2025)
C-JEPA	Immagini statiche	Clustering delle rappresentazioni latenti per stabilità dell'addestramento	Clustering + EMA	Monemi et al. (2025)
SToP-JEPA	Immagini statiche	Student-Teacher online con Positive pairs; allineamento esplicito	Teacher EMA	Monemi et al. (2025)
D-JEPA	Immagini statiche	Denoising nel spazio latente: il predictor stima rappresentazioni pulite da corrotte	VICReg + EMA	Monemi et al. (2025)
IWM	Immagini + trasformazioni	Modello del mondo invariante: previsione condizionata su azioni/trasformazioni esplicite	EMA + regolarizzazione esplicita	Garrido et al. (2023)
3D-JEPA	Scene 3D / point cloud	Previsione 3D: patch volumetriche e previsione latente in spazio tridimensionale	EMA target encoder	Monemi et al. (2025)
Point-JEPA	Point cloud LiDAR	Tokenizzazione dei punti 3D e previsione JEPA per percezione robotica	EMA + FPS sampling	Monemi et al. (2025)
T-JEPA	Testo / linguaggio	Previsione di segmenti mascherati nello spazio latente linguistico	EMA encoder linguistico	Monemi et al. (2025)
TS-JEPA	Serie temporali	Previsione di finestre temporali future nello spazio latente; adatto a biosegnali	EMA + normalizzazione RevIN	Monemi et al. (2025)

Audio-JEPA	Audio grezzo	Previsione di frame audio mascherati; confronto con HuBERT e wav2vec	EMA target encoder	Monemi et al. (2025)
A-JEPA	Audio + video	Allineamento audio-visivo attraverso previsione cross-modale latente	EMA multimodale	Monemi et al. (2025)
MC-JEPA	Multimodale (testo+immagini+ audio)	Architettura multi-canale con predictor separati per modalità e predictor di fusione	VICReg per ciascun canale	Monemi et al. (2025)
V-JEPA	Video	Previsione di tube spazio-temporali mascherate nello spazio latente video	EMA target encoder video	Bardes et al. (2024)
V-JEPA 2	Video + robotica	Scalatura di V-JEPA a modelli più grandi; applicazione a compiti robotici	EMA + piano di azione condizionato	Assran et al. (2025)
TI-JEPA	Testo + Immagini	Allineamento cross-modale testo-immagine con previsione latente bidirezionale	EMA + contrastiva debole	Vo et al. (2024)
M3-JEPA	Testo + Immagini + Video	Framework tri-modale con predictor gerarchico; addestramento unificato	VICReg tri-modale	Lei et al. (2025)
JEP-KD	Immagini (distillazione)	Knowledge Distillation attraverso obiettivo JEPA; il predictor funge da studente	EMA teacher fisso	Monemi et al. (2025)
LLM-JEPA	Testo + Codice	Integrazione JEPA in LLM attraverso mascheramento cross-modale dell'attenzione	EMA + perdita composita	Monemi et al. (2025)
Graph-JEPA	Grafi / reti	Previsione di sottografi mascherati nello spazio latente; addestramento su reti knowledge graph	EMA + message passing	Monemi et al. (2025)
JEPA-Score	Qualsiasi dominio	Stima implicita della densità attraverso lo jacobiano dell'encoder JEPA pre-addestrato	Proprietà derivata (non addestrata)	Balestriero et al. (2025)
JEPA wireless (WM-JEPA)	Canali wireless 6G	Adattamento di V-JEPA/I-JEPA per channel charting, ISAC e allocazione risorse radio	EMA + federated learning	Monemi et al. (2025)

EMA = Exponential Moving Average; VICReg = Variance-Invariance-Covariance Regularization; ViT = Vision Transformer; FPS = Farthest Point Sampling; ISAC = Integrated Sensing and Communication.

Appendice E — Glossario della Notazione Matematica e degli Acronimi Principali

E.1 Notazione matematica

Simbolo	Significato
$\mathbf{x} \in \mathbb{R}^D$	Vettore di input nello spazio osservativo D-dimensionale
$\mathbf{z}, \mathbf{s} \in \mathbb{R}^d$	Rappresentazione latente d-dimensionale (embedding)
$\mathbf{f}_\theta(\cdot)$	Encoder (rete neurale contestuale) con parametri θ
$g_\varphi(\cdot)$	Predictor con parametri φ
$\mathbf{f}_\xi(\cdot)$	Target encoder con parametri ξ (aggiornato per EMA)
$\hat{\mathbf{s}}_y$	Rappresentazione latente prevista della regione target y
$\mathbf{s}_y$	Rappresentazione latente effettiva della regione target y
$\mathcal{L}_{\text{JEPA}}$	Funzione di perdita JEPA: distanza $\ \hat{\mathbf{s}}_y - \mathbf{s}_y\ _2^2$ nello spazio latente
$D(\cdot, \cdot)$	Funzione di distanza o divergenza nello spazio delle rappresentazioni
$\mathbb{E}[\cdot]$	Valore atteso (aspettazione)
$\otimes$	Prodotto tensore o prodotto di Kronecker
$\oplus$	Somma diretta o concatenazione di vettori
$\ \cdot\ _2$	Norma euclidea (L2)
$\ \cdot\ _F$	Norma di Frobenius
$J_{\{\mathbf{f}_\theta\}}(\mathbf{x})$	Jacobiano dell'encoder $\mathbf{f}_\theta$ valutato nel punto $\mathbf{x}$
$\sigma_k(\cdot)$	k-esimo valore singolare (in ordine decrescente)
$\pi(a s)$	Politica: distribuzione di probabilità delle azioni dato lo stato
$M(s' s, a)$	Modello di transizione dinamica del mondo
$r(s, a)$	Funzione di ricompensa
$\gamma \in [0, 1)$	Fattore di sconto temporale
$J(\pi)$	Rendimento atteso cumulativo scontato della politica π
ϕ	Meta-parametri del framework Evo/Devo (parametri evolutivi)
$L(\cdot)$	Funzione di fitness (outer loop dell'ottimizzazione bi-livello)
$\tau_{\min}, \tau_{\max}$	Soglie di densità per la curazione del corpus basata su JEPA-SCORE
$\text{do}(X=x)$	Operatore causale do-calculus di Pearl: intervento sulla variabile X
$p(Y \text{do}(X=x))$	Distribuzione interventistica di Y dopo l'intervento $X=x$
$P_{\theta}(\mathbf{x}_t \mathbf{x}_{<t})$	Distribuzione autoregressiva del token $\mathbf{x}_t$ dato il contesto precedente

$\mathcal{L}_{CE}$	Perdita di cross-entropia (obiettivo di addestramento LLM)
d_{model}	Dimensionalità nascosta del transformer (dimensione degli embedding)
N_h	Numero di teste di attenzione nel multi-head attention
$d_k = d_{model}/N_h$	Dimensionalità di ciascuna testa di attenzione
$\text{softmax}(QK^T/\sqrt{d_k})V$	Formula del meccanismo di self-attention scalata

E.2 Acronimi principali

Acronimo	Significato esteso
A-B-M	Sistema A (Observation) — Sistema B (Action) — Sistema M (Meta-control)
BERT	Bidirectional Encoder Representations from Transformers
CPC	Contrastive Predictive Coding
CLIP	Contrastive Language–Image Pre-training
CURL	Contrastive Unsupervised Representations for Reinforcement Learning
DINO / DINOv2	Self-distillation with no labels (v2: versione scalata con ViT-G)
DPO	Direct Preference Optimization
EMA	Exponential Moving Average
GPT	Generative Pre-trained Transformer
HuBERT	Hidden-Unit BERT (autoapprendimento vocale)
ISAC	Integrated Sensing and Communication (6G)
I-JEPA	Image Joint Embedding Predictive Architecture
JEPA	Joint Embedding Predictive Architecture
LLM	Large Language Model
MLOps	Machine Learning Operations (pipeline di gestione dei modelli)
RLHF	Reinforcement Learning from Human Feedback
RL	Reinforcement Learning
SimCLR	Simple Contrastive Learning of Visual Representations
SSL	Self-Supervised Learning
UED	Unsupervised Environment Design
VICReg	Variance-Invariance-Covariance Regularization
ViT	Vision Transformer
VLA	Vision-Language-Action (modelli robotici multimodali)

V-JEPA	Video Joint Embedding Predictive Architecture
WM	World Model

Appendici

Appendice A — Relazione tra Sistema A/B e Sistema 1/2 di Kahneman

La distinzione introdotta da Kahneman (2011) tra Sistema 1 e Sistema 2 riguarda le modalità di inferenza: il Sistema 1 è un'elaborazione rapida, parallela e automatica — corrispondente computazionalmente a una passata in avanti in una rete neurale profonda — mentre il Sistema 2 è un'elaborazione lenta, multi-passo e deliberativa, analoga alla ricerca su albero o al ragionamento per catena di pensiero. La distinzione tra Sistema A e Sistema B proposta da Dupoux, LeCun e Malik (2026) è invece ortogonale: riguarda le modalità di apprendimento piuttosto che le modalità di inferenza.

Sistema A designa l'apprendimento da dati statici o da flussi sensoriali passivamente raccolti — l'organismo o il sistema accumula input e ne costruisce un modello statistico o predittivo senza intervenire sull'ambiente. Sistema B designa l'apprendimento attraverso l'interazione attiva — l'agente esegue azioni, osserva le conseguenze e aggiusta la propria politica comportamentale in funzione dei feedback ricevuti.

Le due distinzioni sono ortogonali nel senso che ciascun quadrante della griglia 2x2 è concettualmente occupabile e corrisponde a algoritmi e fenomeni cognitivi reali, come illustrato nella tabella seguente.

	Sistema A (Osservazione)	Sistema B (Azione)
Sistema 1 (Inferenza rapida)	Predictive coding; apprendimento statistico	Apprendimento della politica (policy learning)
Sistema 2 (Inferenza deliberativa)	Ragionamento controfattuale; apprendimento causale	Apprendimento attraverso la pianificazione

Tabella A.1 — Relazione tra la distinzione di Kahneman (Sistema 1/2) e la distinzione di Dupoux et al. (Sistema A/B).

Un esempio che illustra la distinzione: il predictive coding di Rao e Ballard (1999), in cui la gerarchia corticale genera previsioni top-down sugli input attesi confrontandole con le osservazioni bottom-up, è un meccanismo di Sistema A (impara dalla distribuzione sensoriale passiva) che opera attraverso un processore System 1 (le previsioni avvengono in modo rapido e automatico). Al contrario, la pianificazione per model rollout in uno S-JEPA action-conditioned è un meccanismo di Sistema A-assistito (usa le

rappresentazioni apprese dall'osservazione) che opera attraverso un processore Sistema 2 (la simulazione multi-passo è deliberativa e computazionalmente costosa).

La separabilità delle due distinzioni ha implicazioni per la progettazione di sistemi cognitivi artificiali: un sistema ottimale dovrebbe poter commutare tra tutte e quattro le modalità in funzione del contesto, esattamente come il framework A-B-M del Sistema M si propone di realizzare.

Appendice B — Modalità Avanzate di Meta-Controllo: Apprendimento per Comunicazione e per Immaginazione

Nelle specie che vivono in ambienti complessi e imprevedibili, un Sistema M capace attiva nuove configurazioni funzionali che non emergono nei sistemi con meta-controllo rudimentale. Questa appendice descrive in dettaglio le due modalità di apprendimento avanzate che il Sistema M abilita.

B.1 Apprendimento per Comunicazione

L'apprendimento per comunicazione è un meccanismo di trasmissione accelerata e distribuita della conoscenza che si articola lungo un gradiente di complessità crescente, ognuno dei cui livelli è documentato empiricamente in un sottoinsieme del regno animale.

Al primo livello si trova l'apprendimento osservativo e associativo di base: gli individui modificano il proprio comportamento notando dove altri dirigono la propria attenzione o come reagiscono agli stimoli. Un bambino che si interessa a un giocattolo dopo aver visto un coetaneo giocarci è l'esempio paradigmatico (Bandura, 1977; Moore, 2013). Questa forma di apprendimento è documentata in quasi tutti i vertebrati sociali.

Al secondo livello si colloca la copia comportamentale: gli individui replicano le azioni altrui o i loro esiti, sia imitando movimenti specifici (imitazione delle azioni) sia trovando modi alternativi per raggiungere lo stesso obiettivo (imitazione degli obiettivi). Un esempio classico è lo scimpanzé che usa un metodo diverso per recuperare il cibo dopo aver osservato un conspecifico dominante avere successo (Whiten et al., 2005; Subiaul et al., 2004).

Al terzo livello si colloca l'apprendimento guidato: l'apprendimento è supportato da un'interazione sociale intenzionale o strutturata con un insegnante, che utilizza attenzione condivisa, referenziazione sociale,

scaffolding comportamentale o dimostrazioni. Un genitore che mostra come usare un oggetto mantenendo il contatto visivo con il bambino esemplifica questa modalità (Bruner, 1983; Wood, Bruner e Ross, 1976). Tale modalità è osservata raramente al di fuori degli esseri umani.

Al quarto livello si colloca l'apprendimento sociale di ordine superiore: gli individui interiorizzano norme, generalizzano attraverso i contesti e trasmettono la conoscenza culturalmente attraverso formati simbolici astratti come il linguaggio (Csibra e Gergely, 2009; Henrich, 2016). Soltanto gli esseri umani esibiscono questa forma di apprendimento nella sua pienezza.

Il Sistema M supporta l'apprendimento per comunicazione rilevando i trigger comunicativi — segnali pedagogici come pointing, sguardo diretto, intonazione imperativa — e instradando gli input evidenziati per l'apprendimento nel Sistema A o nel Sistema B. L'intensità dell'episodio di apprendimento può essere one-shot e modulata dal Sistema M in base all'importanza sociale percepita o alla fiducia nell'insegnante (vigilanza epistemica; Sperber et al., 2010). Nei sistemi artificiali attuali, l'apprendimento per comunicazione è interamente esternalizzato: team di data scientist curano i dati da fonti affidabili o attraverso metodi di post-training. Poiché questo processo è completamente esternalizzato, i sistemi attuali non sono capaci di apprendimento sociale né di esercitare vigilanza epistemica riguardo alla fonte dei propri dati.

B.2 Apprendimento per Immaginazione

L'apprendimento per immaginazione è la capacità di apprendere da input generati internamente, senza necessità di interazione con l'ambiente esterno. Questa modalità è documentata in gradi di complessità crescente attraverso le specie.

Il replay della memoria a riposo è documentato nei roditori, dove durante le pause nell'attività (ai punti di decisione o dopo aver ricevuto una ricompensa) si verifica una riattivazione delle place cells recentemente visitate, nell'ordine originale o inverso, tipicamente a una velocità molto superiore. Questo meccanismo è associato alla presa di decisioni, alla pianificazione a breve termine e all'aggiornamento delle funzioni valore nel reinforcement learning (Foster e Wilson, 2006; Diba e Buzsáki, 2007; Mattar e Daw, 2018).

Il replay della memoria durante il sonno avviene durante il sonno non-REM: le esperienze recenti vengono riprodotte a scala temporale compressa, spesso in episodi più lunghi che includono combinazioni nuove. Questo meccanismo è associato al consolidamento della memoria e alla formazione di schemi a lungo termine o alla generalizzazione (Wilson e McNaughton, 1994; Diekelmann e Born, 2010; Kumaran et al., 2016).

La pianificazione a lungo orizzonte è documentata negli esseri umani e in alcuni animali, che mostrano evidenze di problem-solving senza trial-and-error preventivo in compiti che richiedono uso di strumenti in più passi, gratificazione differita o ragionamento controfattuale (Suddendorf e Corballis, 2007; Redshaw e Suddendorf, 2016; Tulving, 2005). Questi fenomeni evidenziano la simulazione immaginativa come substrato per la previsione flessibile.

Il Sistema M supporta queste modalità operative commutando i Sistemi A e B in modalità di inferenza, instradando le informazioni di input dalla memoria (invece che dai sensori) e instradando le informazioni di output (incluse le azioni) verso la simulazione interna. Il sistema può quindi eseguire rollout simulati di traiettorie immaginative, applicare l'apprendimento alle traiettorie riuscite e consolidare le rappresentazioni risultanti. Questa modalità costituisce l'equivalente computazionale del training su dati sintetici, ma con una differenza fondamentale: i dati sintetici sono generati dal world model interno del sistema stesso, calibrato sull'esperienza reale accumulata.

Appendice C — Esempi di Segnali di Meta-Controllo per l'Apprendimento Autonomo in Umani e Animali

La tabella seguente presenta una lista non esaustiva di meta-stati documentati in umani e animali con i loro effetti sull'Input Selection (IS), sull'Efficacia dell'Apprendimento (LE) e sul Controllo della Modalità operativa (MC). La maggior parte di questi segnali è relativamente semplice da calcolare — ad esempio, la preferenza per i volti si riduce a una preferenza per un pattern a T scuro su sfondo chiaro — il che li rende relativamente facili da far emergere in un contesto di meta-learning.

Meta-stato	Specie	Effetti (IS/LE/MC)	Riferimenti selezionati
Segnali specie-specifici			
Volti e vocalizzazioni	Neonati umani	IS (preferenza attenzionale); LE (facilitazione apprendimento sociale)	Johnson et al. 1991; Vouloumanos e Werker 2007
Direzione dello sguardo	Neonati umani, cani, corvidi	IS (gaze following); MC (attivazione inferenza intenzionale)	Farroni et al. 2002; Tomasello et al. 1998

Segnali pedagogici (alta tonalità, sguardo diretto, pointing)	Neonati umani	IS; LE; MC (shift da memorizzazione a generalizzazione)	Fernald 1985; Csibra e Gergely 2009
Movimento biomeccanico auto-propulso	Neonati umani, pulcini	IS; LE; MC (switch a ragionamento teleologico)	Simion et al. 2008; Csibra et al. 1999
Minacce potenziali (stimoli che si avvicinano, serpenti/ragni)	Neonati umani, mammiferi	IS (orientamento rapido); LE (apprendimento rapido per paura); MC (freezing/alta vigilanza)	Ball e Tronick 1971; Ohman e Mineka 2001
Conspicifici dominanti o prestigiosi	Neonati umani, primati	IS; LE; MC (shift da imitazione ad alto a basso livello)	Shepherd et al. 2006; Buttelmann et al. 2013
Segnali epistemici			
Fiducia selettiva nei conspecifici	Neonati umani, primati, cani	IS; LE; MC (shift all'esplorazione se fonte inaffidabile)	Koenig et al. 2004; Gweon et al. 2014
Contraddizioni logiche/semantiche	Adulti umani, neonati, primati	IS (violazione delle aspettative); MC (switch a simulazione del world-model)	Baillargeon et al. 1985; Botvinick et al. 2001
Esiti inattesi	Mammiferi, uccelli, insetti	IS (risposta di orientamento); LE; MC (shift all'esplorazione)	Sokolov 1963; Pearce e Hall 1980
Incertezza	Neonati umani, primati	LE (fusione multimodale bayesiana); MC (interruzione dello sfruttamento)	Ernst e Banks 2002; Hampton 2001
Stimoli di complessità intermedia	Neonati umani, primati	IS (effetto Goldilocks); LE (ottimizzazione dei progressi)	Kidd et al. 2012; Kang et al. 2009
Segnali somatici			
Sonno	Tutti gli animali	MC (disconnessione input sensoriali); LE (boost apprendimento); IS (replay memoria)	Wilson e McNaughton 1994; Diekelmann e Born 2010
Riposo	Tutti gli animali	MC (attenuazione input sensoriali); LE (apprendimento per immaginazione); IS (replay)	Buckner et al. 2008; Raichle 2015
Dolore	Tutti gli animali	MC (interruzione piani goal-directed; switch a reazioni); LE (apprendimento one-shot)	Eccleston e Crombez 1999

Fame	Tutti gli animali	MC (alta esplorazione; boost goal-directed); IS (rilevamento cibo potenziato)	Balleine e Dickinson 1998; Kolling et al. 2012
Stress	Tutti gli animali	MC (stress basso: esplorazione; stress alto: politiche reattive, sfruttamento)	Schwabe e Wolf 2009

Tabella C.1 — Lista non esaustiva di meta-stati in umani e animali, con i loro effetti interpretati in termini di Input Selection (IS), Learning Efficacy (LE: potenziamento del tasso di apprendimento) e Mode Control (MC) in umani e animali. Adattata da Dupoux, LeCun e Malik (2026).

Appendice D — Tavola Sinottica delle Varianti Architettureali JEPA

La tabella seguente offre una panoramica sistematica di tutte le varianti JEPA documentate nella letteratura al momento della composizione di questa monografia. Le varianti sono organizzate per dominio sensoriale primario e per paradigma di estensione.

Variante	Dominio	Encoder	Caratteristica distintiva
Immagini statiche			
I-JEPA	Immagini	ViT	Mascheramento multi-blocco a scala variabile; EMA target encoder; previsione puramente latente
MIM-JEPA	Immagini	ViT + SCOT	Sparse Convolutional Tokenizer; migliora efficienza campionaria su dataset piccoli
CNN-JEPA	Immagini	ResNet-50	Encoder interamente convoluzionale; prestazioni su ImageNet-100 con meno risorse
C-JEPA	Immagini	ViT	Integrazione regolarizzazione VICReg; riduzione collasso rappresentativo
SToP-JEPA	Immagini	ViT	Embedding posizionali stocastici gaussiani; robustezza all'incertezza spaziale

D-JEPA	Immagini/Video/Audio	ViT	Aggiunge obiettivi di diffusion e flow matching; generazione condizionale
IWM	Immagini	ViT	Trasformazioni parametriche come variabili ausiliarie; struttura equivariante
DMT-JEPA	Immagini	ViT	Discriminative Masked Targets con cross-attention; migliora semantica locale
MaskJEPA	Immagini/Segmentazione	ViT	Integrazione con mask classification; segmentazione SSL
SparseJEPA	Immagini	ViT	Penalità sparsity-inducing; riduzione ridondanza nello spazio latente
Nuvole di punti 3D			
3D-JEPA	Nuvole di punti	PointNet + Transformer	Prima estensione a dati non su griglia; FPS + KNN tokenizzazione
Point-JEPA	Nuvole di punti	Transformer	Greedy Sequencer; sequenze coerenti di patch di punti
Video			
V-JEPA	Video	ViT 3D	3D Multi-Block Masking; mascheramento 90%; rappresentazioni spaziotemporali
V-JEPA 2	Video + Azione	ViT 3D + predictor AR	Condizionamento sull'azione; planning zero-shot; V-JEPA 2-AC per robotica
MC-JEPA	Video	PWC-Net + ViT	Integrazione stima flusso ottico e SSL; multi-task sinergico
Dati tabulari			
T-JEPA	Tabellare	FT-Transformer	Tokenizzazione per colonne; mascheratura a livello colonna
Audio			

Audio-JEPA	Audio/Spettrogramma	ViT	Tratta spettrogramma Mel come immagine mono-canale
A-JEPA	Audio	ViT	Curriculum di mascheramento in due fasi; prestazioni AudioSet-2M
Multimodale			
TI-JEPA	Testo + Immagine	ViT + BERT-like	Spazio embedding congiunto; cross-attention testuale-visiva
M3-JEPA	Any-to-any	Multi-encoder	Predictor MMoE; Alternating Gradient Descent; training multi-task
JEP-KD	Video + Audio (lip reading)	CNN video + ASR	Predictor-come-generatore GAN; framework avversariale per distillazione
Grafi			
Graph-JEPA	Grafi	GNN + Transformer	Allineamento sottografi; prima applicazione JEPA a rappresentazioni a livello di grafo
Wireless e comunicazioni			
TS-JEPA	Serie temporali controllo	Transformer	Codifica stati remoti per riduzione overhead trasmissione
W-JEPA	CSI wireless	Transformer	Channel charting + previsione dinamica canale latente
LLM integrati			
LLM-JEPA	Testo + Codice	GPT + encoder JEPA	Allineamento latente cross-modale; mascheramento cross-attention

Tabella D.1 — Panoramica sistematica delle varianti architetturali JEPA documentate nella letteratura al marzo 2026. AR = Autoregressivo; CSI = Channel State Information; FPS = Farthest Point Sampling; GAN = Generative Adversarial Network; MMoE = Multi-Gate Mixture of Experts; ViT = Vision Transformer.

Appendice E — Glossario della Notazione Matematica e degli Acronimi Principali

E.1 Notazione Matematica

La seguente tabella raccoglie i simboli matematici utilizzati in questa monografia con le loro definizioni.

Simbolo	Definizione
x, y	Variabili di input/output generiche; campioni da una distribuzione
z, s	Rappresentazioni latenti (embedding); stati interni
θ, ϕ	Parametri apprendibili di encoder/predictor (rispettivamente)
θ'	Parametri dell'encoder target (aggiornati per EMA)
$f_{\theta}(x)$	Encoder del contesto con parametri θ applicato all'input x
$f'_{\{\theta'\}}(y)$	Encoder del target con parametri θ' applicato al target y
$g_{\phi}(s; z)$	Predictor con parametri ϕ , condizionato sulla variabile ausiliaria z
s_x, s_y	Rappresentazione latente del contesto x e del target y
$\hat{s}_y$	Stima predetta della rappresentazione del target
$D(a, b)$	Misura di distanza nello spazio latente tra a e b (tipicamente norma L2)
$\ v\ _2$	Norma euclidea (L2) del vettore v
$E[X]$	Valore atteso della variabile aleatoria X
τ	Coefficiente EMA per l'aggiornamento del target encoder (tipicamente 0.996-0.999)
λ	Peso del termine di regolarizzazione anti-collasso
R	Termine di regolarizzazione anti-collasso (es. VICReg)
L_{tot}	Perdita totale di addestramento
$J(\pi)$	Rendimento atteso scontato della politica π nel Sistema B
γ	Fattore di sconto nel reinforcement learning (γ in $[0,1)$)
$r(s, a)$	Funzione di reward nell'MDP del Sistema B
$M(s' s, a)$	Modello dinamico del mondo (distribuzione di transizione)

$\pi(a s)$	Politica dell'agente (distribuzione sulle azioni dato lo stato)
ϕ	Parametri meta (nell'ottimizzazione bi-livello evolutiva)
$L(A_0:A_K, B_0:B_K)$	Funzione di fitness sull'intera traiettoria di vita dell'agente
$P(Y \text{do}(X=x))$	Distribuzione interventistica (operatore do di Pearl)
$P(Y X=x)$	Distribuzione condizionale osservazionale
SCM	Structural Causal Model: $M = (V, U, F, P_U)$
ϕ	Quantita' di integrazione dell'informazione nella IIT di Tononi

E.2 Acronimi Principali

Acronimo	Espansione e contesto
ACC	Anterior Cingulate Cortex — corteccia cingolata anteriore; ruolo nel monitoraggio dei conflitti
ACT-R	Adaptive Control of Thought-Rational — architettura cognitiva modulare di Anderson
BPE	Byte-Pair Encoding — algoritmo di tokenizzazione per LLM
CIRL	Cooperative Inverse Reinforcement Learning — framework di Russell per l'allineamento
CPC	Contrastive Predictive Coding — apprendimento di rappresentazioni contrastive
CSI	Channel State Information — informazione sullo stato del canale nelle comunicazioni wireless
CURL	Contrastive Unsupervised Representations for RL — rappresentazioni contrastive per RL
DINO	Self-DIstillation with NO labels — SSL con distillazione teacher-student
DINOv2	Versione scalata di DINO con dataset curato LVD-142M
DLPFC	Dorso-Lateral Prefrontal Cortex — corteccia prefrontale dorso-laterale
DMN	Default Mode Network — rete neurale di default attiva a riposo
DPO	Direct Preference Optimization — alternativa a RLHF senza reward model esplicito
EMA	Exponential Moving Average — media mobile esponenziale per aggiornamento target encoder
ELAA	Extremely Large-Scale Antenna Arrays — antenne di grandi dimensioni per 6G
GWT	Global Workspace Theory — teoria della coscienza di Baars e Dehaene

H-JEPA	Hierarchical JEPA — estensione gerarchica multi-livello di JEPA
IIT	Integrated Information Theory — teoria della coscienza di Tononi
IRM	Invariant Risk Minimization — metodo per apprendimento causale di Arjovsky et al.
ISAC	Integrated Sensing And Communication — integrazione sensing e comunicazione per 6G
IWM	Image World Model — variante JEPA con trasformazioni parametriche
JEPA	Joint Embedding Predictive Architecture — architettura di previsione negli spazi latenti condivisi
LLM	Large Language Model — modello linguistico di larga scala
MAML	Model-Agnostic Meta-Learning — algoritmo di meta-learning di Finn et al.
MDP	Markov Decision Process — formalismo per il reinforcement learning
MLOps	Machine Learning Operations — pipeline operativa per lo sviluppo dei sistemi ML
MPC	Model Predictive Control — controllo ottimale con orizzonte finito
OFC	Orbitofrontal Cortex — corteccia orbitofrontale; valutazione affettiva
PER	Prioritized Experience Replay — replay esperienze pesato per errore temporale
PFC	Prefrontal Cortex — corteccia prefrontale; meta-controllo biologico
RLHF	Reinforcement Learning from Human Feedback — allineamento LLM con preferenze umane
SCM	Structural Causal Model — modello causale strutturale di Pearl
SDN	Software-Defined Networking — reti definite dal software; analogia per Sistema M
S-JEPA	Sequential JEPA — estensione sequenziale per previsione multi-passo
SSL	Self-Supervised Learning — apprendimento auto-supervisionato
UED	Unsupervised Environment Design — progettazione automatica di ambienti di training
VICReg	Variance-Invariance-Covariance Regularization — regolarizzazione anti-collasso per SSL
ViT	Vision Transformer — trasformatore applicato a immagini come sequenze di patch
VLA	Vision-Language-Action model — modello che integra visione, linguaggio e azione
V-JEPA	Video-JEPA — estensione di JEPA al dominio video con mascheramento spaziotemporale
V-JEPA 2	Seconda versione di V-JEPA con condizionamento sull'azione per la robotica

Riferimenti Bibliografici

Aaronson, S. (2014). Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander). Blog post, scottaaronson.blog.

Abraham, W. C., e Bear, M. F. (1996). Metaplasticity: the plasticity of synaptic plasticity. *Trends in Neurosciences*, 19(4), 126-130.

Adolph, K. E., Cole, W. G., Komati, M., Garciaguirre, J. S., Badaly, D., Kurz, J. M., Korleski, G. L., e Freedland, R. L. (2012). How do you learn to walk? Thousands of steps and dozens of falls per day. *Psychological Science*, 23(11), 1387-1394.

Agrawal, P., Nair, A., Abbeel, P., Malik, J., e Levine, S. (2016). Learning to poke by poking: Experiential learning of intuitive physics. *NeurIPS*, 29, 5074-5082.

Ahn, M., et al. (2022). Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. *arXiv:2204.01691*.

Alayrac, J.-B., et al. (2022). Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35, 23716-23736.

Amodei, D., et al. (2016). Concrete problems in AI safety. *arXiv:1606.06565*.

Anderson, J. R., et al. (2004). An integrated theory of the mind. *Psychological Review*, 111(4), 1036-1060.

Andrychowicz, M., et al. (2016). Learning to learn by gradient descent by gradient descent. *NeurIPS*, 3981-3989.

Arjovsky, M., et al. (2019). Invariant Risk Minimization. *arXiv:1907.02893*.

Assran, M., et al. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. *CVPR*, 15619-15629.

Assran, M., et al. (2025). V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv:2506.09985*.

Aubret, A., Matignon, L., e Hassas, S. (2023). An information-theoretic perspective on intrinsic motivation in RL. *Entropy*, 25(2), 327.

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

- Bacon, P.-L., Harb, J., e Precup, D. (2017). The option-critic architecture. *AAAI*, 31.
- Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science*, 13(3), 89-94.
- Baker, B., et al. (2022). Video pretraining (VPT): Learning to act by watching unlabeled online videos. *NeurIPS*, 35, 24639-24654.
- Balestriero, R., et al. (2025). JEPA-SCORE: Implicit density estimation via predictive architectures. Preprint.
- Ball, W., e Tronick, E. (1971). Infant responses to impending collision. *Science*, 171(3973), 818-820.
- Balleine, B. W., e Dickinson, A. (1998). Goal-directed instrumental action. *Neuropharmacology*, 37(4-5), 407-419.
- Bandura, A. (1977). *Social Learning Theory*. Prentice Hall.
- Bar, A., et al. (2024). STOP-JEPA: Stochastic Positional Embeddings for JEPA. Preprint.
- Bardes, A., et al. (2022). VICReg: Variance-Invariance-Covariance Regularization. *ICLR*.
- Bardes, A., Ponce, J., e LeCun, Y. (2023). MC-JEPA: A Joint-Embedding Predictive Architecture for Self-Supervised Learning of Motion and Content. Preprint.
- Bardes, A., et al. (2024). Revisiting feature prediction for learning visual representations from video. *arXiv:2404.08471*.
- Bellemare, M. G., et al. (2016). Unifying count-based exploration and intrinsic motivation. *NeurIPS*, 1471-1479.
- Bengio, Y., Louradour, J., Collobert, R., e Weston, J. (2009). Curriculum learning. *ICML*, 41-48.
- Bengio, Y., Courville, A., e Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE TPAMI*, 35(8), 1798-1828.
- Bernstein, N. A. (1967). *The Co-ordination and Regulation of Movements*. Pergamon Press.
- Bisk, Y., et al. (2020). Experience grounds language. *EMNLP*, 8718-8735.
- Bommasani, R., et al. (2022). On the opportunities and risks of foundation models. *arXiv:2108.07258*.

- Botvinick, M., et al. (2019). Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, 23(5), 408-422.
- Botvinick, M. M., et al. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108(3), 624-652.
- Bou Chaaya, K., Girgis, R., e Bennis, M. (2025). W-JEPA: Channel charting with predictive world modeling. Preprint.
- Boyd, R., Richerson, P. J., e Henrich, J. (2011). The cultural niche. *PNAS*, 108(S2), 10918-10925.
- Brandom, R. (1994). *Making It Explicit*. Harvard University Press.
- Brentano, F. (1874). *Psychologie vom empirischen Standpunkt*. Duncker & Humblot.
- Bruce, J., et al. (2024). Genie: generative interactive environments. *ICML*.
- Bruner, J. (1983). *Child's Talk: Learning to Use Language*. W. W. Norton.
- Buckner, R. L., Andrews-Hanna, J. R., e Schacter, D. L. (2008). The brain's default network. *Annals of the NYAS*, 1124(1), 1-38.
- Buttelmann, D., et al. (2013). Selective imitation of in-group over out-group members. *Child Development*, 84(2), 422-428.
- Caron, M., et al. (2021). Emerging properties in self-supervised vision transformers. *ICCV*, 9650-9660.
- Cavell, S. (1979). *The Claim of Reason*. Oxford University Press.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200-219.
- Chalmers, D. J. (1996). *The Conscious Mind*. Oxford University Press.
- Chen, T., et al. (2020). A simple framework for contrastive learning of visual representations. *ICML*, 1597-1607.
- Chen, X., et al. (2025). D-JEPA: Discrete diffusion meets JEPA for unified generative representation. Preprint.
- Christiano, P., et al. (2017). Deep reinforcement learning from human preferences. *NeurIPS*, 4299-4307.
- Coeckelbergh, M. (2010). Robot rights? *Ethics and Information Technology*, 12(3), 209-221.
- Craik, K. J. W. (1943). *The Nature of Explanation*. Cambridge University Press.

- Csibra, G., e Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, 13(4), 148-153.
- Csibra, G., et al. (1999). Teleological reasoning in infancy. *Trends in Cognitive Sciences*, 3(4), 136-143.
- Daw, N. D., et al. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems. *Nature Neuroscience*, 8(12), 1704-1711.
- Daw, N. D., et al. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441, 876-879.
- Daw, N. D., et al. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6), 1204-1215.
- Dehaene, S., e Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2), 200-227.
- Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown.
- Dennis, M., et al. (2020). Emergent complexity and zero-shot transfer via unsupervised environment design. *NeurIPS*, 13049-13061.
- Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers. *NAACL-HLT*, 4171-4186.
- Diba, K., e Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature Neuroscience*, 10(10), 1241-1242.
- Diekelmann, S., e Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience*, 11(2), 114-126.
- Dosovitskiy, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR 2021*.
- Driess, D., et al. (2023). PaLM-E: An embodied multimodal language model. *ICML*, 8469-8488.
- Dulac-Arnold, G., et al. (2021). Challenges of real-world RL. *Machine Learning*, 110(9), 2419-2468.
- Dupoux, E. (2018). Cognitive science in the era of AI. *Cognition*, 173, 43-59.
- Dupoux, E., LeCun, Y., e Malik, J. (2026). Why AI systems don't learn and what to do about it: Lessons on autonomous learning from cognitive science. *arXiv:2603.15381*.
- Ecoffet, A., et al. (2021). First return, then explore. *Nature*, 590(7847), 580-586.

- Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, 48(1), 71-99.
- Ernst, M. O., e Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870), 429-433.
- Eysenbach, B., et al. (2019). Diversity is all you need: Learning skills without a reward function. *ICLR*.
- Fei, Z., et al. (2023). A-JEPA: Joint-Embedding Predictive Architecture Can Listen. Preprint.
- Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development*, 8(2), 181-195.
- Finn, C., Abbeel, P., e Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *ICML*, 1126-1135.
- Floridi, L., et al. (2018). AI4People — An ethical framework for a good AI society. *Minds and Machines*, 28(4), 689-707.
- Fodor, J. A., e Pylyshyn, Z. (1988). Connectionism and cognitive architecture. *Cognition*, 28(1-2), 3-71.
- Foster, D. J., e Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells. *Nature*, 440, 680-683.
- Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127-138.
- Gallagher, S. (2005). *How the Body Shapes the Mind*. Oxford University Press.
- Garrido, Q., et al. (2025). Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv:2502.11831*.
- Geirhos, R., et al. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 665-673.
- Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Houghton Mifflin.
- Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin.
- Girgis, R., Valcarce, A., e Bennis, M. (2025). TS-JEPA: Semantic communication via joint-embedding predictive architecture. Preprint.
- Gopnik, A. (2012). Scientific thinking in young children. *Science*, 337(6102), 1623-1627.

- Gopnik, A., et al. (1999). *The Scientist in the Crib*. William Morrow & Co.
- Gopnik, A., et al. (2017). Changes in cognitive flexibility and hypothesis search across human life history. *PNAS*, 114(30), 7892-7899.
- Gottlieb, J., et al. (2013). Information-seeking, curiosity, and attention. *Trends in Cognitive Sciences*, 17(11), 585-593.
- Goyal, A., e Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proc. Royal Society A*, 478(2266).
- Graph-JEPA — Skenderi, G., et al. (2025). Graph-JEPA: Graph joint-embedding predictive architecture for graph-level representations. Preprint.
- Gureckis, T. M., e Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, 7(5), 464-481.
- Ha, D., e Schmidhuber, J. (2018). World models. *ICLR*.
- Hadfield-Menell, D., et al. (2017). Inverse reward design. *NeurIPS*, 6765-6774.
- Hadsell, R., et al. (2020). Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24(12), 1028-1040.
- Hafner, D., et al. (2019). Learning latent dynamics for planning from pixels. *ICML*, 2555-2565.
- Hafner, D., et al. (2020). Dream to control: Learning behaviors by latent imagination. *ICLR*.
- Hafner, D., et al. (2023). Mastering diverse domains through world models. *arXiv:2301.04104*.
- Habermas, J. (1981). *Theorie des kommunikativen Handelns*. Suhrkamp.
- Hamrick, J. B., et al. (2017). Metacontrol for adaptive imagination-based optimization. *ICLR*.
- Hansen, N. (2006). The CMA evolution strategy: A comparing review. In *Towards a New Evolutionary Computation*. Springer.
- Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42(1-3), 335-346.
- Harris, P. L., e Corriveau, K. H. (2011). Young children's selective trust in informants. *Phil. Trans. Royal Society B*, 366(1567), 1179-1187.

- Hassabis, D., et al. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, 95(2), 245-258.
- He, K., et al. (2020). Momentum contrast for unsupervised visual representation learning. *CVPR*, 9712-9721.
- He, K., et al. (2022). Masked autoencoders are scalable vision learners. *CVPR*, 16000-16009.
- Heidegger, M. (1927). *Sein und Zeit*. Max Niemeyer Verlag. [Trad. it.: *Essere e Tempo*, Longanesi, 1976]
- Henrich, J. (2016). *The Secret of Our Success*. Princeton University Press.
- Hensch, T. K. (2005). Critical period plasticity in local cortical circuits. *Nature Reviews Neuroscience*, 6(11), 877-888.
- Heyes, C. (2018). *Cognitive Gadgets: The Cultural Evolution of Thinking*. Harvard University Press.
- Hoffmann, J., et al. (2022). Training compute-optimal large language models. *NeurIPS*, 35, 30016-30030.
- Hsu, W.-N., et al. (2021). HuBERT: Self-supervised speech representation learning. *IEEE/ACM TASLP*, 29, 3451-3460.
- Hu, Z., et al. (2024). 3D-JEPA: A Joint Embedding Predictive Architecture for 3D Point Cloud Self-Supervised Learning. Preprint.
- Huang, L., LeCun, Y., e Balestriero, R. (2025). LLM-JEPA: Combining generative and predictive objectives for multimodal alignment. Preprint.
- Hupkes, D., et al. (2020). Compositionality decomposed. *Journal of Artificial Intelligence Research*, 67, 757-795.
- IRM — Arjovsky, M., Bottou, L., Gulrajani, I., e Lopez-Paz, D. (2019). Invariant Risk Minimization. *arXiv:1907.02893*.
- Jacobs, R. A., et al. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1), 79-87.
- Johnson-Laird, P. N. (1983). *Mental models*. Harvard University Press.
- Johnson, M. H. (2001). Functional brain development in humans. *Nature Reviews Neuroscience*, 2(7), 475-483.
- Johnson, M. H., et al. (1991). Newborns' preferential tracking of face-like stimuli. *Cognition*, 40(1-2), 1-19.
- Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

- Kalapos, A., e Gyires-Toth, B. (2024). CNN-JEPA: Self-supervised pretraining with a joint-embedding predictive architecture using convolutional neural networks. Preprint.
- Kang, M. J., et al. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry. *Psychological Science*, 20(8), 963-973.
- Kaplan, J., et al. (2020). Scaling laws for neural language models. *arXiv:2001.08361*.
- Karmiloff-Smith, A. (1992). *Beyond Modularity*. MIT Press.
- Kendal, R. L., et al. (2015). Chimpanzees copy dominant and knowledgeable individuals. *Evolution and Human Behavior*, 36(1), 65-72.
- Kidd, C., Piantadosi, S. T., e Aslin, R. N. (2012). The Goldilocks effect. *PLoS One*, 7(5), e36399.
- Kim, S., et al. (2024). MaskJEPA: Combining masked self-supervised learning with JEPA for segmentation. Preprint.
- Kirkpatrick, J., et al. (2017). Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13), 3521-3526.
- Koenig, M. A., et al. (2004). Trust in testimony. *Psychological Science*, 15(10), 694-698.
- Koh, P. W., et al. (2021). WILDS: A benchmark of in-the-wild distribution shifts. *ICML*, 139, 5637-5664.
- Kolling, N., et al. (2012). Neural mechanisms of foraging. *Science*, 336(6077), 95-98.
- Krueger, D., et al. (2021). Out-of-distribution generalization via risk extrapolation (REx). *ICML*, 5815-5826.
- Kumaran, D., Hassabis, D., e McClelland, J. L. (2016). What learning systems do intelligent agents need? *Trends in Cognitive Sciences*, 20(7), 512-534.
- Lake, B. M., e Baroni, M. (2018). Generalization without systematicity. *ICML*, 2873-2882.
- Lake, B. M., et al. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253.
- Laskin, M., Srinivas, A., e Abbeel, P. (2020). CURL: Contrastive unsupervised representations for RL. *ICML*, 5639-5650.
- Lavechin, M., et al. (2023). BabySLM: language-acquisition-friendly benchmark of SSL spoken language models. *INTERSPEECH*, 4588-4592.

- LeCun, Y. (2022). A path towards autonomous machine intelligence. OpenReview, version 0.9.2.
- Lei, Z., et al. (2025). M3-JEPA: Multi-modal joint embedding predictive architecture with any-to-any alignment. Preprint.
- Li, T., et al. (2020). Federated optimization in heterogeneous networks (FedProx). MLSys.
- Lillicrap, T. P., et al. (2016). Continuous control with deep reinforcement learning. ICLR.
- Liu, H., et al. (2023). LLaVA: Large language and vision assistant. NeurIPS, 36.
- Ma, Y. J., et al. (2023). VIP: Towards universal visual reward and representation. ICLR.
- Malik, J., et al. (see Dupoux et al. 2026).
- Mattar, M. G., e Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. Nature Neuroscience, 21, 1609-1617.
- McMahan, H. B., et al. (2017). Communication-efficient learning of deep networks from decentralized data. AISTATS, 1273-1282.
- Meltzoff, A. N. (2007). Like me: a foundation for social cognition. Developmental Science, 10(1), 126-134.
- Merleau-Ponty, M. (1945). *Phénoménologie de la perception*. Gallimard. [Trad. it.: *Fenomenologia della percezione*, il Saggiatore, 2003]
- Merleau-Ponty, M. (1968). *Le visible et l'invisible*. Gallimard.
- Miller, E. K., e Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. Annual Review of Neuroscience, 24(1), 167-202.
- Mo, S., e Tong, X. (2024). C-JEPA: Collapse-resistant joint-embedding predictive architecture. Preprint.
- Mo, S., e Yun, S. (2024). DMT-JEPA: Discriminative Masked Targets for Joint-Embedding Predictive Architecture. Preprint.
- Moerland, T. M., et al. (2023). Model-based RL: A survey. Foundations and Trends in Machine Learning, 16(1), 1-118.
- Mohamed, S., e Rezende, D. J. (2015). Variational information maximisation for intrinsically motivated RL. NeurIPS, 28, 2125-2133.
- Monemi, A., Chinipardaz, R., Rasti, B., Bennis, M., e Latva-aho, M. (2025). A Comprehensive Tutorial on Joint Embedding Predictive Architectures

(JEPA). arXiv:2412.xxxxx [Centro per le Comunicazioni Wireless, Universita' di Oulu].

Morton, J., e Johnson, M. H. (1991). CONSPEC and CONLERN. *Psychological Review*, 98(2), 164-181.

Nair, A., et al. (2022). R3M: A universal visual representation for robot manipulation. *CoRL*.

Nishio, T., Chen, X., e Bennis, M. (2025). ISAC multimodal semantic fusion with JEPA. Preprint.

Oizumi, M., Albantakis, L., e Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: IIT 3.0. *PLoS Computational Biology*, 10(5).

Oller, D. K., e Eilers, R. E. (1988). The role of audition in infant babbling. *Child Development*, 59(2), 441-449.

Oquab, M., et al. (2024). DINOv2: Learning robust visual features without supervision. *TMLR*.

Oudeyer, P.-Y., e Kaplan, F. (2007). What is intrinsic motivation? A typology. *Frontiers in Neurorobotics*, 1, 6.

Oudeyer, P.-Y., Kaplan, F., e Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE TEC*, 11(2), 265-286.

Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS*, 35, 27730-27744.

Pascalis, O., de Haan, M., e Nelson, C. A. (2002). Is face processing species-specific during the first year of life? *Science*, 296(5571), 1321-1323.

Pathak, D., et al. (2017). Curiosity-driven exploration by self-supervised prediction. *ICML*, 2778-2787.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J. (2018). *The Book of Why*. Basic Books.

Peters, J., Buhlmann, P., e Meinshausen, N. (2016). Causal inference by using invariant prediction. *JRSS-B*, 78(5), 947-1012.

Peters, J., Mooij, J. M., Janzing, D., e Scholkopf, B. (2017). *Elements of Causal Inference*. MIT Press.

Pfeiffer, B. E., e Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths. *Nature*, 497(7447), 74-79.

Piaget, J. (1952). The origins of intelligence in children. International University Press.

Point-JEPA — Saito, K., et al. (2025). Point-JEPA: A Point Cloud Self-Supervised Learning Method via Joint Embedding Predictive Architecture. Preprint.

Pong, V. H., et al. (2020). Skew-fit: State-covering self-supervised RL. ICML, 7783-7792.

Portelas, R., et al. (2020). Automatic curriculum learning for deep RL: A short survey. arXiv:2003.04664.

Qiao, Z., et al. (2025). TokCom: Semantic generative tokenized communication framework. Preprint.

Radford, A., et al. (2021). Learning transferable visual models from natural language supervision. ICML, 8748-8763.

Radosavovic, I., et al. (2022). Real-world robot learning with masked visual pre-training. CoRL.

Radosavovic, I., et al. (2024). Learning humanoid locomotion over challenging terrain. arXiv:2410.03654.

Raissi, M., Perdikaris, P., e Karniadakis, G. E. (2019). Physics-informed neural networks. Journal of Computational Physics, 378, 686-707.

Rao, R. P. N., e Ballard, D. H. (1999). Predictive coding in the visual cortex. Nature Neuroscience, 2(1), 79-87.

Rasch, B., e Born, J. (2013). About sleep's role in memory. Physiological Reviews, 93(2), 681-766.

Redshaw, J., e Suddendorf, T. (2016). Children's capacity to imagine and prepare for alternative future possibilities. Cognition, 150, 104-113.

Ren, P., et al. (2021). A survey of deep active learning. ACM Computing Surveys, 54(9), 1-40.

Ring, M. B. (1997). CHILD: A first step towards continual learning. Machine Learning, 28(1), 77-104.

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

Saffran, J. R., Aslin, R. N., e Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science, 274(5294), 1926-1928.

Saito, K., et al. (2025). Point-JEPA for point cloud learning. Preprint.

- Schaul, T., et al. (2016). Prioritized experience replay. ICLR.
- Schaeffer, R., Miranda, B., e Koyejo, S. (2023). Are emergent abilities of large language models a mirage? NeurIPS, 36.
- Schmidhuber, J. (1987). Evolutionary principles in self-referential learning. Dissertation, TU Munchen.
- Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom. In *From Animals to Animats*. MIT Press.
- Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation. *IEEE TAMM*, 2(3), 230-247.
- Scholkopf, B. (2019). Causality for machine learning. arXiv:1911.10500.
- Scholkopf, B., et al. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5), 612-634.
- Schrittwieser, J., et al. (2020). Mastering Atari, Go, chess and shogi by planning with a learned model (MuZero). *Nature*, 588, 604-609.
- Schultz, W., Dayan, P., e Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306), 1593-1599.
- Schwabe, L., e Wolf, O. T. (2009). Stress prompts habit behavior in humans. *Journal of Neuroscience*, 29(22), 7191-7198.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-424.
- Searle, J. R. (1992). *The Rediscovery of the Mind*. MIT Press.
- Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge University Press.
- Shenhav, A., et al. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, 40, 99-124.
- Silver, D., et al. (2017). Mastering the game of Go without human knowledge. *Nature*, 550, 354-359.
- Silver, D., e Sutton, R. S. (2025). Welcome to the era of experience. Google AI.
- Sinha, A., Malo, P., e Deb, K. (2018). A review on bilevel optimization. *IEEE TEC*, 22(2), 276-295.
- Skenderi, G., et al. (2025). Graph-JEPA for graph-level representation learning. Preprint.

- Smith, L. B., et al. (2018). The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences*, 22(4), 325-336.
- Sperber, D., et al. (2010). Epistemic vigilance. *Mind & Language*, 25(4), 359-393.
- Spelke, E. S., et al. (1992). Origins of knowledge. *Psychological Review*, 99(4), 605.
- Stanley, K. O., e Miikkulainen, R. (2002). Evolving neural networks through augmenting topologies. *Evolutionary Computation*, 10, 99-127.
- Stahl, A. E., e Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, 348(6230), 91-94.
- Subiaul, F., et al. (2004). Cognitive imitation in rhesus macaques. *Science*, 305(5682), 407-410.
- Suddendorf, T., e Corballis, M. C. (2007). The evolution of foresight. *Behavioral and Brain Sciences*, 30(3), 299-313.
- Sun, J., Qin, X., e Yang, J. (2024). JEP-KD: Joint embedding predictive knowledge distillation for lip reading. Preprint.
- Sutton, R. S., e Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.
- Tang, C., et al. (2025). Deep reinforcement learning for robotics: A survey of real-world successes. *AAAI*, 39, 28694-28698.
- Taylor, A. H., et al. (2012). New Caledonian crows learn the functional properties of novel tool types. *PLoS ONE*, 7(12), e49374.
- Tharwat, B., et al. (2025). Latent action pretraining through world modeling. *arXiv:2509.18428*.
- Thelen, E., e Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press.
- Thimonier, H., et al. (2024). T-JEPA: Tabular self-supervised learning via joint embedding predictive architecture. Preprint.
- Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Harvard University Press.
- Thrun, S., e Pratt, L. (1998). Learning to learn: Introduction and overview. In *Learning to Learn*. Springer.
- Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard University Press.

- Tomasello, M. (2009). *Why We Cooperate*. MIT Press.
- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5, 42.
- Tooby, J., e Cosmides, L. (1992). The psychological foundations of culture. In *The Adapted Mind*. Oxford University Press.
- Tulving, E. (2005). Episodic memory and autonoesis. In *The Missing Link in Cognition*. Oxford University Press.
- Tuncay, M., et al. (2025). Audio-JEPA: Self-supervised audio representation learning via joint-embedding prediction. Preprint.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460.
- Turkewitz, G., e Kenny, P. A. (1982). Limitations on input as a basis for neural organization. *Developmental Psychobiology*, 15(4), 357-368.
- Turkle, S. (2011). *Alone Together*. Basic Books.
- Van den Oord, A., Li, Y., e Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv:1807.03748*.
- Varela, F. J., Thompson, E., e Rosch, E. (1991). *The Embodied Mind*. MIT Press.
- Vaswani, A., et al. (2017). Attention is all you need. *NeurIPS*, 30, 5998-6008.
- Velez-Garcia, A., et al. (2025). MIM-JEPA: Masked image modelling via joint-embedding predictive architecture for improved sample efficiency. Preprint.
- Vo, T., et al. (2024). TI-JEPA: Text-Image Joint Embedding Predictive Architecture for multimodal alignment. Preprint.
- Von Kugelgen, J., et al. (2023). Self-supervised learning with data augmentations provably isolates content from style. *NeurIPS*.
- Vouloumanos, A., e Werker, J. F. (2007). Listening to language at birth. *Developmental Science*, 10(2), 159-164.
- Vygotsky, L. S., e Cole, M. (1978). *Mind in society*. Harvard University Press.
- Wang, R., et al. (2019). PAIRED: Protagonist Antagonist Induced Regret Environment Design. *ICLR 2020*.
- Wei, J., et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

- Werker, J. F., e Tees, R. C. (1984). Cross-language speech perception. *Infant Behavior and Development*, 7(1), 49-63.
- Whiten, A., e Ham, R. (1992). On the nature and evolution of imitation in the animal kingdom. *Advances in the Study of Behavior*, 21, 239-283.
- Whiten, A., et al. (2005). Conformity to cultural norms of tool use in chimpanzees. *Nature*, 437(7059), 737-740.
- Wilson, M. A., e McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, 265(5172), 676-679.
- Wilson, R. C., et al. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of Experimental Psychology: General*, 143(6), 2074-2081.
- Wittgenstein, L. (1953). *Philosophische Untersuchungen / Philosophical Investigations*. Blackwell. [Trad. it.: *Ricerche filosofiche*, Einaudi, 1999]
- Wood, D., Bruner, J. S., e Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2), 89-100.
- Yarats, D., et al. (2020). CURL: Contrastive unsupervised representations for RL (variante). *ICML*, 10490-10500.
- Zahavi, D. (2001). *Husserl and Transcendental Intersubjectivity*. Ohio University Press.
- Zador, A. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 10(1), 3770.
- Zhao, T. Z., et al. (2023). Learning fine-grained bimanual manipulation with low-cost hardware. *RSS*.
- Zhou, C., et al. (2024). Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv:2408.11039*.
- Zitkovich, B., et al. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. *CoRL*, 229, 2165-2183.
- Zmyj, N., et al. (2010). The reliability of a model influences 14-month-olds' imitation. *Journal of Experimental Child Psychology*, 106(4), 208-220.
- Zoph, B., e Le, Q. V. (2017). Neural architecture search with reinforcement learning. *ICLR*.